6-14-2018

Phylogeny, Taxonomy, and the Concept of Lineage in the Presence of Horizontal Gene Transfer

Matthew Fullmer
University of Connecticut - Storrs, matthew.fullmer@uconn.edu

Phylogeny, Taxonomy, and the Concept of Lineage in the Presence of Horizontal Gene Transfer

Matthew S. Fullmer
University of Connecticut, 2018

Whole genome sequencing has opened enormous worlds of opportunity in recent years as the number of sequenced organisms has continued to skyrocket. Keeping track of what we have sequenced and added to our databases, as the cornucopia of data grows ever larger, is essential. Maintaining order in our classification of genomes and understanding how they relate to each other is probable to only continue to grow in importance as time progresses. A second major consequence of the explosion in genome sequencing is the ever-increasing opportunity to explore the distribution and role of rare genes in the pan-genomes of phylogenetic groups and communities as never before. Such insights offer us opportunities to glean how these genes at the seeming periphery of a group can direct organismal interactions and perhaps shape or repress the emergence of new lineages.

The first section of this thesis discusses how established methodologies can elucidate both phylogeny and taxonomy. Tools such as multi-locus sequence analysis, average nucleotide identity (ANI), and core genome phylogenies are shown to converge on the same answers to how genomes relate to one another and how they should be classified. The second section discusses a novel extension to the ANI concept. This extension allows the inference of statistically supported phylogenies from whole genome data. As a byproduct of this new method deeper taxonomic ranks can now be delimited by *in silico* genomic distance. A

detection and identification pipeline for restriction-methylation systems in the class Halobacteria is presented in the third section. Additionally, the strong proclivity of these genes to be transferred across the breadth of the class is also analyzed. Finally, the last section discusses a hypothesis of how apparently mutualistic interactions could arise through a process of mutual cheating. Furthermore, the hypothesis is compared with prior hypotheses that also invoke distributed genomes and shared functions.

Phylogeny, Taxonomy, and the Concept of Lineage in the Presence of Horizontal Gene Transfer

Matthew S. Fullmer

B.S., University of Massachusetts, Amherst, 2007

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2018

APPROVAL PAGE

Doctor of Philosophy Dissertation

Phylogeny, Taxonomy, and the Concept of Lineage in
the Presence of Horizontal Gene Transfer

Presented by
Matthew S. Fullmer, B.S.

Major Advisor

_____

J. Peter Gogarten

Associate Advisor

_____

R. Thane Papke

Associate Advisor

_____

Joerg Graf

Associate Advisor

_____

Spencer Nyholm

Associate Advisor

_____

Jonathan Klassen

Associate Advisor

_____

Paul O. Lewis

University of Connecticut
2018

Acknowledgements

I wish to thank, first and foremost, J. Peter Gogarten. His patience and encouragement have been enormously important to me over the years. I cannot imagine my time at UConn having gone any better under any one else.

Several members of the Gogarten lab require specific mention. Erica Lasek-Nesselquist, a post-doc in our lab during my early years, was a vital mentor to me. Shannon Soucy, a fellow student in the lab, was a valued friend and colleague and provided an excellent outlet for both casual banter and serious science discussion. Sean Gosselin developed from a part-time undergraduate to an independant researcher before my eyes. He is equally adept at discussing history as he is at evolution. Artemis Louyakis has provided great advice on the terminal stages of a Ph.D. and has helped keep me on point while writing.

My committee, R. Thane Papke, Joerg Graf, Spencer Nyholm, Jonathan Klassen, and Paul Lewis, have all made important marks on both my work. Even more importatnly, they have greatly influenced how I approach and conduct science, particulalry in the collaborative environments that are now so common.

Finally, I would not be the kind of person who could tackle a Ph.D. program without the support of my family. My parents, Steven and Marie, instilled in me many of the same traits that I have found most valuable in academia. My sisters, Sarah and Margaret, have shown me nothing but love and support which has been crucial to maintaining my morale. My wife Amanda, it almost goes without saying, has been paramount in my successes here. Her encouragement to go back to school, understanding during low moments, and the occasional kicks to the rear have all been essential.

Table of Contents

Chapter 1 - General Introduction and Outline of Chapters

Chapter 1.1 A brief history of taxonomy

Taxonomy is the science of classification of living things (Boone and Castenholz, 2012). The modern concept of taxonomy begins with Linnaeus and moves forward from there (Cain, 1957). Linnaeus placed the genus as the unit of import with species far less important and defined by a long series of up to twelve *differentiae* (Cain, 1958). In fact, the well-known single word species names in the binomial system were mere afterthoughts in his system and not intended for serious use. These important genera were the result of applying the Aristotelian principle of "Logical Division" to create "Natural Groupings (Cain, 1962)." These principles strove to define groups by only their "essences." Any other attributes were viewed as "accidental" attributes to be discarded for purpose of classification (Cain, 1957, 1962).

The nature of how to divide organisms into these logical groupings is at the heart of taxonomy. While geometry and other mathematics offer relatively clean examples of essential versus accidental attributes, such as the definition of a circle as opposed to the color of the line used to draw it, the biological world is rarely so forthcoming. Biological classification has a long history of needing to change its mind on what it values as essential attributes (Cain, 1962). For a long time emphasis was placed on using *a priori* principles that were then applied to organisms (Cain, 1958, 1962). However, these systems created situations where it was patently obvious that things that were related to each other were separated or that different organisms were being grouped together, such

as using the wood production group monocots and dicots together (Cain, 1962; Cowan, 1962). Rather, so-called "blind groping," perhaps fairly summarized as human intuition, was creating the groupings that we then sought to define and quantify through rules (Cain, 1957, 1962). When systems were deemed to be succeeding, albeit usually temporarily, it was because they agreed with the natural groups our intuitions had already largely settled on.

By the time of Darwin the preferred *a priori* principles were organ systems deemed essential (Cain, 1958). However, these were fraught with difficulties over cases such as undeveloped and vestigial organs that did not fit into the contemporary schema. Darwin recognized and criticized these approaches on the grounds that one inevitably creates absurd groupings, such as the varying development of eyes (Darwin, 1859). His theory of evolution provided a useful explanation. Taxonomists could explain their models' difficulty with attributes such as vestigial organs through a history of shared descent (Cain, 1962). Taxonomy was taking its first steps into phylogeny (de Queiroz, 1996). An important consequence was a shifting emphasis to non-overlapping, nested, and mutually exclusive groups (de Queiroz, 1996; Hennig, 1965). Darwin's theory, however, did still create an obstacle for taxonomy. A danger lay in assuming that similar attributes indicated a shared heritage. Thus, any system of classification that is based on evolution must distinguish convergence from ancestral resemblance. Nevertheless, taxonomists frequently treated shared descent as a superficial *post hoc* explanation and justification of the natural groups they constructed (de Queiroz, 1996).

By the middle of the 20th century taxonomists could generally be split into two schools of thought (Sokal, 1963). The first were those considering phylogenetic origins of

biological attributes in their classification schema, and the second were those who used only comparative evidence in the absence of phylogenetic consideration (Sokal, 1963). The later, sometimes calling themselves empirical taxonomists, argued that classification was meaningless without declaring the purpose of the exercise and that such purposes should be practical (Blackwelder, 1967; Sokal, 1963). These empiricists began turning to statistics to improve help build their groups through the "new method" of numerical taxonomy (Cain, 1962; Sokal, 1963). Numerical taxonomy is a strictly phenetic system, eschewing any consideration of phylogenetics from its considerations. Looking back at it from today, one wonders if it was a form of backlash to what de Quiroz (1996) termed "the evolutionizing of taxonomy." In any case, the methods frequently found results highly congruent with the accepted taxonomies (Blackwelder, 1967). Yet again, the tacit ultimate barometer of our efforts is what we innately intuit to be so. Numerical taxonomy claimed to be a leap forward by forgoing the *a priori* weighting of characters in classical methods and replacing those with a belief that all characters are equivalent (Blackwelder, 1967). This assumption was challenged on multiple grounds (Blackwelder, 1967). First, assuming all characters are equal is assigning a weight, however agnostic it may be. Second, the proponents admitted that the method requires the use of "taxonomically informative" characters, which belies the idea of equality (Blackwelder, 1967; Sokal, 1963).

Meanwhile, the phylogenetic camp continued "evolutionizing" their ideas. Hennig was a prime mover, championing "phylogenetic systematics." In this system, all relationships are judged through the lens of vertical descent with an emphasis on attaining strict monophyly of groups (Hennig, 1965) defined by shared derived characters

(synapomorphies). Any characters that have either remained unchanged for long periods of time (plesiomorphies) or have only recently emerged (autapomorphies) should be eschewed, because they would over-unite or over-divide groups in phenetic systems (Hennig, 1965). This new approach to classification was paralleled by a developing shift from conceiving of species as groups of similar individuals into populations of interbreeding individuals (de Queiroz, 1996).

Molecular phylogenetics began in earnest during the 1960s and began influencing taxonomic thinking (Zuckerkandl and Pauling, 1965). Research focused primarily on comparing protein sequences and inferring phylogenies from parts thereof (Fitch and Margoliash, 1967). Although, nucleic acids were already appreciated as stores of molecular history data (Zuckerkandl and Pauling, 1965). However, the real leap forward in emphasizing genealogy over phenetics in microbes came with the introduction of SSU rRNAs as phylogenetic chronometers (Fox et al., 1977; Woese and Fox, 1977). These developments, first catalogues of small oligomer libraries and then later partial and full gene sequencing, generated a push in genealogy towards "natural taxonomies" or "natural classification" systems (Pace et al., 2012; Sapp, 2007; Woese et al., 1990). Natural taxonomic systems were not new to the bacteria and had actually been proposed decades earlier by Stanier, van Neil, and Kluyver (Kluyver et al., 1936; Stanier and Niel, 1941). Interestingly, Stanier and van Neil felt evidence for their systems was both lacking and unlikely to be forthcoming and expressed an unwillingness to defend them in later work (Stanier and Niel, 1962). These newer rRNA-based systems were predominantly concerned with organizing organisms by the principles of shared descent. Interestingly, this shift may have been the first time that an *a priori* principle created major changes in

classification and was not promptly torn down for destroying established groups. However, it is worth noting that microbiologists have generally had fewer characters to work with and less time to accrue expectations and preconceptions of how microbes should group than botanists and zoologists (Cain, 1962; Cowan, 1962). Regardless, the phylogenetic emphasis gained steam and informed debate on classification, taxonomy, and nomenclature (Sapp, 2007). The methods used have migrated into other, and multiples of, housekeeping genes (Naser et al., 2005; Sullivan et al., 2005) as researchers sought the ability to resolve more closely related taxa, and as limitations to the use of solely 16S sequencing became apparent (Boucher et al., 2004; Morandi et al., 2005). These new methods, and 16S phylogenetics before them, all seem to carry the torch from the earliest days of taxonomy. They all seek to discern which genes reflect the essential attributes of the organism's history of vertical descent. Old schemes are discarded because some or all of the genes are found to suffer incomplete lineage sorting, horizontal gene transfer, or some other malady that interferes with detection of shared descent.

Somewhat in parallel to the rRNA and sequencing-driven revolution, another line of taxonomic analysis developed (Johnson, 1985; Steigerwalt et al., 1976). DNA-DNA Hybridization (DDH) is based on comparing the amount of hybridization between two genomes (Steigerwalt et al., 1976). High values indicate that the query comparator shares large amounts of very similar genomic DNA to the reference. Below a certain threshold, usually 70%, the two comparators can be ruled as not belonging to the same species. DDH eventually emerged as an accepted "gold standard" for delimiting species (Stackebrandt and Goebel, 1994). One of the most interesting aspects to this approach from an historical taxonomic viewpoint is how it shares traits with the earlier phonetic

and classical taxonomy principles. It is explicitly called a phylogenetic method (Stackebrandt and Goebel, 1994), but there is no attempt made to verify that all of the hybridizing DNA shares a common history. The measurement also looks at the bulk, or net, signal of the organism. If one imagines each base pair as a individual character then one might be forgiven for seeing a parallel to the numerical taxonomy concept in the calculation. The success of DDH has prompted many researchers to attempt to augment or supplant the method with comparisons of whole genome sequences (Auch et al., 2010a; Goris et al., 2007; Konstantinidis and Tiedje, 2005; Meier-Kolthoff et al., 2013; Richter and Rosselló-Móra, 2009; Varghese et al., 2015). It might be argued that these systems are a throwback as much to the pre-phylogenetic ideas of taxonomy as they are to advancing the natural taxonomy principles of phylogenetic taxonomy.

Chapter 1.2 What do we want from species?

A recurring theme throughout the evolution of taxonomy is questioning what we wish our systems of classification and nomenclature to represent (Blackwelder, 1967; Cain, 1962; Cowan, 1962; de Queiroz, 1996; Sokal, 1963; Woese et al., 1990; Woese and Fox, 1977). This is not an idle point. At the heart of many disputes lies a fundamental disagreement over taxonomy's priority (Blackwelder, 1967; Sapp, 2007; Sokal, 1963; Woese et al., 1990). This conflict is perhaps quite natural. After all, the users of our taxonomic schema come from different backgrounds and have different uses for them. While an evolutionary biologist may have good cause to favor a system that quickly and efficiently organizes taxa by history of shared descent, a clinician may have equally sound reasons for desiring a classification that groups taxa by pathogenic potential and

modes of action in humans over other considerations. Ultimately, these tensions come to a head at the species level.

The mere word species is a loaded one. Where once Linnaeus saw only the genus as the rank of prime importance (Cain, 1958), it has atrophied and the species has slowly rose to supplant it entirely (Cain, 1962; de Queiroz, 1996). Indeed, some researchers view species as the only taxonomic rank that is worth trying to define (Stackebrandt and Goebel, 1994). Others doubt even its very existence (Cowan, 1962; Lawrence, 2002; Papke et al., 2007). Its long history of definitions and redefinitions has led to some researchers coining new terms to avoid its baggage (Fullmer et al., 2014b; Papke et al., 2007) or struggle with how HGT impacts species concepts and shared ancestry (Dykhuizen and Green, 1991; Gogarten et al., 2002).

The classical taxonomical approach, to put like with like, suggests that species need not require its members to be closely related. The phenetic school of numerical taxonomy certainly would not. The phylogentic school that has gained prominence places a history of shared descent at the apex of its priorities and assumes incongruous characters are the equivalent of the old accidental attributes. As the phylogenetic philosophy is currently ascendant it is tempting to call it the winner and accept it as the way to proceed. However, there remain some important points to consider.

The phylogenetic school of taxonomy assumes we can identify the components of the genome that reflect an organism or taxon's history of shared descent. It assumes that we can identify the essential attributes while successfully screening out the accidental ones. This may be a very dangerous assumption. We saw with rRNA phylogenetics that its genes do not necessarily reflect a perfect history of vertical transmission (Boucher et

al., 2004). Additionally, what is the correct course of action when an organism possesses multiple divergent copies of its rRNA gene (Morandi et al., 2005)? Adding more genes will not necessarily solve the problem (Salichos and Rokas, 2013). From the microbiologist's perspective, the core problem is horizontal gene transfer (HGT). HGT allows genes to break their linkage with the genome and each other. This means these genes can experience different histories and convolute phylogenetic analysis (Colston et al., 2014; Papke et al., 2004). The rate of HGT has been reported at high levels in many groups (Khomyakova et al., 2011; Olendzenski et al., 2000; Sharma et al., 2007; Williams et al., 2011; Zhaxybayeva et al., 2006, 2009). This does not only disrupt individual gene phylogenies. It may be large enough to affect the signals measured in whole genome comparisons. Methods such as DDH (Steigerwalt et al., 1976), ANI (Goris et al., 2007; Konstantinidis and Tiedje, 2005), and GGDC (Auch et al., 2010a) must tacitly assume, if they are to be viewed as phylogenetic, that the shared signal is vertical descent rather than the result of some other process. Otherwise, they are perhaps best viewed as being of a phenetic nature in line with numerical taxonomy (Sokal, 1963). Worryingly for the phylogenetic school, large effective rates of HGT are able to mimic the patterns in genomic relatedness associated with vertical descent (Andam and Gogarten, 2011). As HGT is more frequent between closely related organisms (Fraser et al., 2007; Williams et al., 2012) it can be assumed that many clusterings we observe are vertical descent augmented by biased gene transfer. Yet, cases like the *Thermotoga* (Zhaxybayeva et al., 2009), where the genome appears to be a patchwork of at least three disparate sources, should give any researcher pause before making that assumption without justification.

So where does that leave us? Certainly, our ability to accurately classify organisms into groups by shared common descent is far from certain at this point, although the amount of uncertainty is different at different taxonomic levels. We can classify on highly practical traits, such as many clinicians prefer (speaking anecdotally, this author is personally familiar with multiple clinicians who use *Shigella* as functional definition for any microbe that produces shiga-toxins). But these schemata may not be of much value outside of their narrow designed-for scope. Perhaps those who throw up their hands at the entire affair are on to something (Cowan, 1962; Papke et al., 2007)? But then why, even when we are skeptical of taxonomy, do we return to it (in this author's case, see chapter 3)? The only conclusion that seems certain is that scientists will continue attempting to concoct a perfect system that addresses all of the difficulties.

Chapter 1.3 Chapter Overviews

One of the prevalent themes of my work is identifying and delimiting species, or similar units, and what has shaped and continues to shape these units. To the former, I have used well-established classification methods to successfully contribute to resolving multiple taxonomic groups. Additionally, I have contributed some novel extensions and additions to the field's cadre of tools. To the latter, I have looked at natural populations and communities in the halophilic Archaea to investigate how mating barriers may be constructed to reduce homogenization of communal lineages. I also have played significant parts in research aiming to understand the unit of horizontal transfer in the genus *Aeromonas*. Extending beyond the direct research, I have spent time thinking about

how populations might be coaxed into different evolutionary paths and what the results might look like.

The second chapter of this work focuses on the application and incremental extension of several well-established or increasingly well-established methodologies for taxon delineation. The primary players here are multi-locus sequencing analysis (MLSA), average nucleotide identity (ANI), and *in silico* DNA-DNA hybridization (*is*DDH or dDDH) (Auch et al., 2010a; Goris et al., 2007; Konstantinidis and Tiedje, 2005; Sullivan et al., 2005). All three were developed before I became a graduate student, and all three were on the way up as I began my work. MLSA was in a stage of incrementally adding more genes to the analysis in the hopes of strengthening the results by tamping down on the discord of horizontal gene transfer (HGT). I took advantage of my access to collaborations with large amounts of whole genome data and ramped the number of genes up from as few as two or three up to 5 (Fullmer et al., 2014b), 16 (Colston et al., 2014), and even past thirty (Collins et al., 2015; Gromek et al., 2016). I also took the method to, or arguably past, its logical conclusion and made expanded-core phylogenies from all of the common genes in a dataset (Colston et al., 2014). While initially optimistic about the power of MLSA and the stories it could tell, I was under no illusions about the frailty of relying on a single line of evidence. As a result, I became increasingly interested in whole-genome comparisons. This concept most commonly takes the form of ANI (Konstantinidis and Tiedje, 2005; Richter and Rosselló-Móra, 2009), wherein the entirely of an assembled genome is compared against the entirety of a second assembled genome. I found this method dovetailed very well with the MLSA results and allowed me to make

strong conclusions about the relationships of many of the taxa in my datasets. I added an *is*DDH method (Auch et al., 2010b) to the *Aeromonas* study. While conceptually very similar to ANI it uses a different algorithm and adds post-processing to place its result on the same scale as traditional wet-lab DDH values, allowing for direct comparisons. This method offered several advantages over ANI, such as statistical measures of uncertainty around the point estimates, and a more intrinsic incorporation of the fractions of genomes used in its calculations. Its results were largely in close concordance with the other methods, although it was discovered that we needed to innovate slightly to adapt the method to the *Aeromonas* genus.

My collaborators and I introduced new MLSA schemata, participated at the leading edge of increasing their scope, adapted *is*DDH to meet the demands of a convoluted genus, and applied some of these methodologies to groups that had seen little to no exposure to them prior, as well as on new scales. However, the methods used were still largely by-the-book or only modestly innovative. The third chapter covers my exploration into more uncharted territory with the development of a new iteration of ANI and the implicit discovery of potentially informative cutoffs for deeper taxonomic ranks than traditional ANI or *is*DDH have been able to offer.

The fourth chapter covers my foray into seeking out and examining rare genes in the Halobacteria. I start by discussing the role and our understanding of HGT in the Halobacteria (Fullmer et al., 2014a). I then move on to discussing my progress in identifying restriction-methylation genes in communities of the genus *Halorubrum* and

the Halobacteria class as a whole. Evidence is presented for the methylase genes undergoing transfer to the point of being freely available and speculation on how methylation may play less of a role in divergence than I assumed.

The final chapter veers in a different direction with a philosophical treatise on the nature of pan-genomes. I present a notion where the pan-genome acts a shared resource for some or all members of a species, population, or community (Fullmer et al., 2015). The arrival at this hypothesized state, seemingly of pure cooperation, might actually be the result of purely cheating by all participants.

References

Andam, C. P., and Gogarten, J. P. (2011). Biased gene transfer and its implications for the concept of lineage. *Biol. Direct* 6, 47. doi:10.1186/1745-6150-6-47.

Auch, A. F., Jan, M. von, Klenk, H.-P., and Göker, M. (2010a). Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand. Genomic Sci.* 2, 117–134. doi:10.4056/sigs.531120.

Auch, A. F., Klenk, H.-P., and Göker, M. (2010b). Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. *Stand. Genomic Sci.* 2, 142–148. doi:10.4056/sigs.541628.

Blackwelder, R. E. (1967). A Critique of Numerical Taxonomy. *Syst. Zool.* 16, 64. doi:10.2307/2411518.

Boone, D. R., and Castenholz, R. W. (2012). *Bergey's Manual of Systematic Bacteriology: Volume One : The Archaea and the Deeply Branching and Phototrophic Bacteria*. Springer Science & Business Media.

Boucher, Y., Douady, C. J., Sharma, A. K., Kamekura, M., and Doolittle, W. F. (2004). Intragenomic heterogeneity and intergenomic recombination among haloarchaeal rRNA genes. *J. Bacteriol.* 186, 3980–3990. doi:10.1128/JB.186.12.3980-3990.2004.

Cain, A. J. (1957). Logic and Memory in Linnaeus's System of Taxonomy. *Proc. Linn. Soc. Lond.* 169, 144–163. doi:10.1111/j.1095-8312.1958.tb00819.x.

Cain, A. J. (1958). The Post-Linnaean Development of Taxonomy. *Proc. Linn. Soc. Lond.* 170, 234–244. doi:10.1111/j.1095-8312.1959.tb00857.x.

Cain, A. J. (1962). "The Evolution of Taxonomic Principles," in *Microbial classification.* Available at: https://www.cabdirect.org/cabdirect/abstract/19621102214 [Accessed June 7, 2018].

Collins, A. J., Fullmer, M. S., Gogarten, J. P., and Nyholm, S. V. (2015). Comparative genomics of Roseobacter clade bacteria isolated from the accessory nidamental gland of Euprymna scolopes. *Front. Microbiol.* 6. doi:10.3389/fmicb.2015.00123.

Colston, S. M., Fullmer, M. S., Beka, L., Lamy, B., Gogarten, J. P., and Graf, J. (2014). Bioinformatic Genome Comparisons for Taxonomic and Phylogenetic Assignments Using Aeromonas as a Test Case. *mBio* 5, e02136-14. doi:10.1128/mBio.02136-14.

Cowan, S. T. (1962). "The Microbial Species - A Macromyth," in *Microbial classification.* Available at: https://www.cabdirect.org/cabdirect/abstract/19621102214 [Accessed June 7, 2018].

Darwin, C. (1859). *On the Origin of Species*. London: Routledge.

de Queiroz, K. (1996). The Linnaean Hierarchy and the Evolutionization of Taxonomy, with Emphasis on the Problem of Nomenclature. *Aliso* 15, 125–144. doi:10.5642/aliso.19961502.07.

Dykhuizen, D. E., and Green, L. (1991). Recombination in Escherichia coli and the definition of biological species. *J. Bacteriol.* 173, 7257–7268. doi:10.1128/jb.173.22.7257-7268.1991.

Fitch, W. M., and Margoliash, E. (1967). Construction of Phylogenetic Trees. *Science* 155, 279–284. doi:10.1126/science.155.3760.279.

Fox, G. E., Magrum, L. J., Balch, W. E., Wolfe, R. S., and Woese, C. R. (1977). Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proc. Natl. Acad. Sci.* 74, 4537–4541. doi:10.1073/pnas.74.10.4537.

Fraser, C., Hanage, W. P., and Spratt, B. G. (2007). Recombination and the nature of bacterial speciation. *Science* 315, 476–480. doi:10.1126/science.1127573.

Fullmer, M. S., Gogarten, J. P., and Papke, R. T. (2014a). "Horizontal Gene Transfer in Halobacteria," in *Halophiles: Genetics and Genomes* (Caister Academic Press), 196.

Fullmer, M. S., Soucy, S. M., and Gogarten, J. P. (2015). The pan-genome as a shared genomic resource: mutual cheating, cooperation and the black queen hypothesis. *Evol. Genomic Microbiol.*, 728. doi:10.3389/fmicb.2015.00728.

Fullmer, M. S., Soucy, S. M., Swithers, K. S., Makkay, A. M., Wheeler, R., Ventosa, A., et al. (2014b). Population and genomic analysis of the genus Halorubrum. *Extreme Microbiol.* 5, 140. doi:10.3389/fmicb.2014.00140.

Gogarten, J. P., Doolittle, W. F., and Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* 19, 2226–2238.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., and Tiedje, J. M. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* 57, 81–91. doi:10.1099/ijs.0.64483-0.

Gromek, S. M., Suria, A. M., Fullmer, M. S., Garcia, J. L., Gogarten, J. P., Nyholm, S. V., et al. (2016). Leisingera sp. JC1, a Bacterial Isolate from Hawaiian Bobtail Squid Eggs, Produces Indigoidine and Differentially Inhibits Vibrios. *Front. Microbiol.* 7. doi:10.3389/fmicb.2016.01342.

Hennig, W. (1965). Phylogenetic Systematics. *Annu. Rev. Entomol.* 10, 97–116.

Johnson, J. L. (1985). "2 DNA Reassociation and RNA Hybridisation of Bacterial Nucleic Acids," in *Methods in Microbiology*, ed. T. Bergan (Academic Press), 33–74. doi:10.1016/S0580-9517(08)70471-9.

Khomyakova, M., Bükmez, Ö., Thomas, L. K., Erb, T. J., and Berg, I. A. (2011). A Methylaspartate cycle in haloarchaea. *Science* 331, 334–337. doi:10.1126/science.1196544.

Kluyver, A. J., Niel, V., and B, C. (1936). Prospects for a Natural System of Classification of Bacteria. *Zentralblatt Bakteriol. Parasitenkd. Infekt. Hyg.* 94, 369–403.

Konstantinidis, K. T., and Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 102, 2567–2572. doi:10.1073/pnas.0409727102.

Lawrence, J. G. (2002). Gene transfer in Bacteria: speciation without species? *Theor. Popul. Biol.* 61, 449–460. doi:10.1006/tpbi.2002.1587.

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14, 60. doi:10.1186/1471-2105-14-60.

Morandi, A., Zhaxybayeva, O., Gogarten, J. P., and Graf, J. (2005). Evolutionary and diagnostic implications of intragenomic heterogeneity in the 16S rRNA gene in Aeromonas strains. *J. Bacteriol.* 187, 6561–6564. doi:10.1128/JB.187.18.6561-6564.2005.

Naser, S. M., Thompson, F. L., Hoste, B., Gevers, D., Dawyndt, P., Vancanneyt, M., et al. (2005). Application of multilocus sequence analysis (MLSA) for rapid identification of Enterococcus species based on rpoA and pheS genes. *Microbiology* 151, 2141–2150. doi:10.1099/mic.0.27840-0.

Olendzenski, L., Liurid="*," L., Zhaxybayeva, O., Murpheyrid="&Dagger, R., Shin, D.-G., et al. (2000). Horizontal transfer of archaeal genes into the Deinococcaceae: detection by molecular and computer-based approaches. *J. Mol. Evol.* 51, 587–599. doi:10.1007/s002390010122.

Pace, N. R., Sapp, J., and Goldenfeld, N. (2012). Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. *Proc. Natl. Acad. Sci.* 109, 1011–1018. doi:10.1073/pnas.1109716109.

Papke, R. T., Koenig, J. E., Rodríguez-Valera, F., and Doolittle, W. F. (2004). Frequent recombination in a saltern population of halorubrum. *Science* 306, 1928–1929. doi:10.1126/science.1103289.

Papke, R. T., Zhaxybayeva, O., Feil, E. J., Sommerfeld, K., Muise, D., and Doolittle, W. F. (2007). Searching for species in haloarchaea. *Proc. Natl. Acad. Sci.* 104, 14092–14097. doi:10.1073/pnas.0706358104.

Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci.* 106, 19126–19131. doi:10.1073/pnas.0906412106.

Salichos, L., and Rokas, A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497, 327–331. doi:10.1038/nature12130.

Sapp, J. (2007). The structure of microbial evolutionary theory. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* 38, 780–795. doi:10.1016/j.shpsc.2007.09.011.

Sharma, A. K., Walsh, D. A., Bapteste, E., Rodriguez-Valera, F., Doolittle, W. F., and Papke, R. T. (2007). Evolution of rhodopsin ion pumps in haloarchaea. *BMC Evol. Biol.* 7, 79. doi:10.1186/1471-2148-7-79.

Sokal, R. R. (1963). The Principles and Practice of Numerical Taxonomy. *Taxon* 12, 190. doi:10.2307/1217562.

Stackebrandt, E., and Goebel, B. M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *Int. J. Syst. Bacteriol.* 44, 846–849. doi:10.1099/00207713-44-4-846.

Stanier, R. Y., and Niel, C. B. V. (1941). The main outlines of bacterial classification. *J Bacteriol*, 437–466.

Stanier, R. Y., and Niel, C. B. van (1962). The concept of a bacterium. *Arch. Für Mikrobiol.* 42, 17–35. doi:10.1007/BF00425185.

Steigerwalt, A. G., Fanning, G. R., Fife-Asbury, M. A., and Brenner, D. J. (1976). DNA relatedness among species of Enterobacter and Serratia. *Can. J. Microbiol.* 22, 121–137. doi:10.1139/m76-018.

Sullivan, C. B., Diggle, M. A., and Clarke, S. C. (2005). Multilocus sequence typing. *Mol. Biotechnol.* 29, 245–254. doi:10.1385/MB:29:3:245.

Varghese, N. J., Mukherjee, S., Ivanova, N., Konstantinidis, K. T., Mavrommatis, K., Kyrpides, N. C., et al. (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Res.* 43, 6761–6771. doi:10.1093/nar/gkv657.

Williams, D., Fournier, G. P., Lapierre, P., Swithers, K. S., Green, A. G., Andam, C. P., et al. (2011). A Rooted Net of Life. *Biol. Direct* 6, 45. doi:10.1186/1745-6150-6-45.

Williams, D., Gogarten, J. P., and Papke, R. T. (2012). Quantifying homologous replacement of loci between haloarchaeal species. *Genome Biol. Evol.* 4, 1223–1244. doi:10.1093/gbe/evs098.

Woese, C. R., and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci.* 74, 5088–5090. doi:10.1073/pnas.74.11.5088.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci.* 87, 4576–4579. doi:10.1073/pnas.87.12.4576.

Zhaxybayeva, O., Gogarten, J. P., Charlebois, R. L., Doolittle, W. F., and Papke, R. T. (2006). Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events. *Genome Res.* 16, 1099–1108. doi:10.1101/gr.5322306.

Zhaxybayeva, O., Swithers, K. S., Lapierre, P., Fournier, G. P., Bickhart, D. M., DeBoy, R. T., et al. (2009). On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. *Proc. Natl. Acad. Sci.* 106, 5865–5870. doi:10.1073/pnas.0901260106.

Zuckerkandl, E., and Pauling, L. (1965). Molecules as documents of evolutionary history. *J. Theor. Biol.* 8, 357–366. doi:10.1016/0022-5193(65)90083-4.

Chapter 2: Molecular phylogenies can discriminate phylogeny and taxonomy

Multi-Locus Sequence Analysis (MLSA; also known as MLST – for Typing) has been an established mechanism for constructing phylogenies of closely related taxa and assisting in classifying species for decades (Zeigler, 2003). It evolved originally out of a need to find an alternative to partial 16S SSU rDNA gene phylogenetics. Sequencing the 16S suffers from several drawbacks in these situations. Firstly, 16S's great advantage in large-scale phylogenetics, its extreme level of conservation, is a liability at a genus, species, or strain level (Hanage et al., 2005). Even its variable regions have had little opportunity to evolve informative mutations among the recently or incompletely diverged taxa. Second, it is a single gene. Any horizontal transfer, partial or whole, can poison any "correct" signal the gene possesses. An assumption, exemplified by the complexity hypothesis (Jain et al., 1999), has held that the sheer number of intracellular interactions 16S rRNAs are involved in would impair the rate of successful transfers to other organisms. The analogy could be made that the 16S genes are like the engines in cars; most of them are highly similar, but the connections, gearing interfaces, and attachment points are rarely the same from model-to-model, compromising the safety or effective operation of a vehicle with an alternative engine. Unfortunately, this has turned out to be not the case with many ribosomal RNAs and proteins (Boucher et al., 2004). The high level of conservation actually allows these ORFs to be transferred more easily because more of the interaction details are preserved (Wang and Zhang, 2000). Finally, a problem of more

limited, but very serious (where occurring), scope is the existence of multiple heterogeneous copies of 16S rRNAs (Gogarten et al., 2002; Lopez-Lopez et al., 2007; Morandi et al., 2005). In the case of the *Aeromonas* the heterogeneity is great enough to change the species assignment in many strains (Morandi et al., 2005). This phenomenon presents several problems. If confronted with multiple divergent 16S ORFs, which should one pick? What happens if the sequencing generates chimeras where the final molecule tells the story of none of its constituents? What if the existing genomic sequences are already chimeras on account of internal recombination? All of these issues combined to create an environment where new alternatives were needed to generate useful phylogenies.

Research with some organisms had begun to move to single-copy housekeeping genes to supplement or replace their 16S phylogenies (Gevers et al., 2005; Naser et al., 2005; Sullivan et al., 2005). These genes had some clear advantages over 16S. Most obviously, they were less conserved which allowed for more resolution at low taxonomic levels. Additionally, being single-copy the issue of multiple heterogeneous copies was ameliorated. However, the other limitations still existed. Any HGT in a dataset could corrupt the result to the point of uselessness for the taxa affected. And these genes were often still so conserved that they could easily move from one organism to another without too much difficulty. The levels of conservation were also often still so high that there could be few-if-any informative sites in alignments (Salichos and Rokas, 2013).

The next leap was to combine multiple single-copy housekeeping genes. The premise was that by using two or more genes, typically concatenated into a single "super gene," the weakly voiced stories from each would combine into a chorus singing a tale of the organism's past. Besides increased resolution there was also the advantage that the (presumed) infrequent transfer events would be washed out by the story as a whole.

Average Nucleotide Identity (ANI) was first developed in 2005 (Konstantinidis and Tiedje, 2005) to compare taxa at a whole-genome level. It was envisioned as a complement or replacement to the DNA-DNA Hybridization (DDH) procedure, which was fraught with varying methodologies, repeatability, and scalability issues. ANI underwent several revisions before a popular version (Richter and Rosselló-Móra, 2009) with a highly accessible GUI (these undoubtedly go hand-in-hand) emerged as an attempt to shift the "gold-standard" for species identification and classification. Meanwhile, another bioinformatic whole-genome comparison, the Genome-to-Genome Distance Calculator (GGDC; I often refer to it as *is*DDH), tool had been developed in parallel and was also reaching functional maturity around the same time (Auch et al., 2010b, 2010a; Meier-Kolthoff et al., 2013). The GGDC also aimed to complement or supplant DDH, but had the advantages of scaling its results directly onto the DDH scale rather than finding a single equivalence point. Additionally, the method includes internal confidence statistics of its estimates. I will cover more on the details and merits of ANI and *is*DDH in Chapter 3.

The work presented in this chapter includes considerable focus on utilizing these tools to classify organisms. The first publication (Fullmer et al., 2014) is a population-based analysis of *Halorubrum* genomes isolated from an Iranian salt lake. It uses an established MLSA scheme as well as ANI to identify the relationships between the isolates and reference taxa. The second section features a second first-author publication (Colston, Fullmer et al., 2014) examining taxonomic and phylogenetic assignments with bioinformatic comparisons. This manuscript uses a novel MLSA scheme, ANI, *is*DDH, and a core genome MLSA to probe the misclassifications and phylogeny of the *Aeromonas*. Additionally, the appendices include several manuscripts where I used these tools in supporting roles to others' research. These include: i) An MLSA providing phylogenetic context to RAPD genomic fingerprinting (Ram Mohan et al., 2014). ii) An ANI and MLSA providing classification and phylogenetic positioning for alpha-proteobacterial isolates of the *Roseobacter* clade (Collins et al., 2015). iii) The phylogenetic placement and several genome comparisons to relatives of an *E. scalopes* egg jelly coat isolate (Gromek et al., 2016).

References

Auch, A. F., Jan, M. von, Klenk, H.-P., and Göker, M. (2010a). Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand. Genomic Sci.* 2, 117–134. doi:10.4056/sigs.531120.

Auch, A. F., Klenk, H.-P., and Göker, M. (2010b). Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. *Stand. Genomic Sci.* 2, 142–148. doi:10.4056/sigs.541628.

Boucher, Y., Douady, C. J., Sharma, A. K., Kamekura, M., and Doolittle, W. F. (2004). Intragenomic heterogeneity and intergenomic recombination among haloarchaeal rRNA genes. *J. Bacteriol.* 186, 3980–3990. doi:10.1128/JB.186.12.3980-3990.2004.

Collins, A. J., Fullmer, M. S., Gogarten, J. P., and Nyholm, S. V. (2015). Comparative genomics of Roseobacter clade bacteria isolated from the accessory nidamental gland of Euprymna scolopes. *Front. Microbiol.* 6. doi:10.3389/fmicb.2015.00123.

Fullmer, M. S., Soucy, S. M., Swithers, K. S., Makkay, A. M., Wheeler, R., Ventosa, A., et al. (2014). Population and genomic analysis of the genus Halorubrum. *Extreme Microbiol.* 5, 140. doi:10.3389/fmicb.2014.00140.

Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., et al. (2005). Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.* 3, 733–739. doi:10.1038/nrmicro1236.

Gogarten, J. P., Doolittle, W. F., and Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* 19, 2226–2238.

Gromek, S. M., Suria, A. M., Fullmer, M. S., Garcia, J. L., Gogarten, J. P., Nyholm, S. V., et al. (2016). Leisingera sp. JC1, a Bacterial Isolate from Hawaiian Bobtail Squid Eggs, Produces Indigoidine and Differentially Inhibits Vibrios. *Front. Microbiol.* 7. doi:10.3389/fmicb.2016.01342.

Hanage, W. P., Fraser, C., and Spratt, B. G. (2005). Fuzzy species among recombinogenic bacteria. *BMC Biol.* 3, 6. doi:10.1186/1741-7007-3-6.

Jain, R., Rivera, M. C., and Lake, J. A. (1999). Horizontal gene transfer among genomes: The complexity hypothesis. *Proc. Natl. Acad. Sci.* 96, 3801–3806. doi:10.1073/pnas.96.7.3801.

Konstantinidis, K. T., and Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 102, 2567–2572. doi:10.1073/pnas.0409727102.

Lopez-Lopez, A., Benlloch, S., Bonfa, M., Rodriguez-Valera, F., and Mira, A. (2007). Intragenomic 16S rDNA divergence in Haloarcula marismortui is an adaptation to different temperatures. *J. Mol. Evol.* 65, 687–696. doi:10.1007/s00239-007-9047-3.

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14, 60. doi:10.1186/1471-2105-14-60.

Morandi, A., Zhaxybayeva, O., Gogarten, J. P., and Graf, J. (2005). Evolutionary and diagnostic implications of intragenomic heterogeneity in the 16S rRNA gene in Aeromonas strains. *J. Bacteriol.* 187, 6561–6564. doi:10.1128/JB.187.18.6561-6564.2005.

Naser, S. M., Thompson, F. L., Hoste, B., Gevers, D., Dawyndt, P., Vancanneyt, M., et al. (2005). Application of multilocus sequence analysis (MLSA) for rapid identification of Enterococcus species based on rpoA and pheS genes. *Microbiology* 151, 2141–2150. doi:10.1099/mic.0.27840-0.

Ram Mohan, N., Fullmer, M. S., Makkay, A. M., Wheeler, R. W., Ventosa, A., Naor, A., et al. (2014). Evidence from phylogenetic and genome fingerprinting analyses suggests rapidly changing variation in Halorubrum and Haloarcula populations. *Extreme Microbiol.* 5, 143. doi:10.3389/fmicb.2014.00143.

Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci.* 106, 19126–19131. doi:10.1073/pnas.0906412106.

Salichos, L., and Rokas, A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497, 327–331. doi:10.1038/nature12130.

Sullivan, C. B., Diggle, M. A., and Clarke, S. C. (2005). Multilocus sequence typing. *Mol. Biotechnol.* 29, 245–254. doi:10.1385/MB:29:3:245.

Wang, Y., and Zhang, Z. (2000). Comparative sequence analyses reveal frequent occurrence of short segments containing an abnormally high number of non-random base variations in bacterial rRNA genes. *Microbiology* 146, 2845–2854. doi:10.1099/00221287-146-11-2845.

Zeigler, D. R. (2003). Gene sequences useful for predicting relatedness of whole genomes in bacteria. *Int. J. Syst. Evol. Microbiol.* 53, 1893–1900. doi:10.1099/ijs.0.02713-0.

Chapter 2.1 Population and genomic analysis of the genus *Halorubrum*

This section consists of both my first 1st author publication, and also my first peer-reviewed publication in (Fullmer et al., 2014b). The major findings of this article include an investigation of the *Halorubrum* populations existing in a presumed island population dynamic. The study reports the relationships of the genomic isolates using both gene-based and whole-genome methods. Additionally, the molecular parasite distributions were examined and apparent barriers to gene flow were inferred. Matthew S. Fullmer, J. Peter Gogarten, Antonio Ventosa, and R. Thane Papke participated in the design of this study and helped to draft the manuscript. Shannon M. Soucy generated the intein data and performed the majority of the intein analysis and helped to draft the manuscript. Kristen S. Swithers performed the CRISPR Recognition Tool analysis and helped to draft the manuscript. Andrea M. Makkay and Ryan Wheeler performed the MLSA PCR. Andrea M. Makkay performed the genome sequencing. All authors read and approved the final manuscript.

frontiers in MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 11 April 2014
doi: 10.3389/fmicb.2014.00140

Population and genomic analysis of the genus *Halorubrum*

Matthew S. Fullmer[1], Shannon M. Soucy[1], Kristen S. Swithers[1,2], Andrea M. Makkay[1], Ryan Wheeler[1], Antonio Ventosa[3], J. Peter Gogarten[1] and R. Thane Papke[1]*

[1] Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA
[2] Department of Cell Biology, Yale School of Medicine, Yale University, New Haven, CT, USA
[3] Department of Microbiology and Parasitology, University of Seville, Seville, Spain

Edited by:
Jesse Dillon, California State University, Long Beach, USA

Reviewed by:
Jesse Dillon, California State University, Long Beach, USA
Federico Lauro, University of New South Wales, Australia

***Correspondence:**
R. Thane Papke, Microbiology Program, Department of Molecular and Cell Biology, University of Connecticut, 91 N. Eagleville Rd., Storrs, CT 06269-3125, USA
e-mail: thane@uconn.edu

The Halobacteria are known to engage in frequent gene transfer and homologous recombination. For stably diverged lineages to persist some checks on the rate of between lineage recombination must exist. We surveyed a group of isolates from the Aran-Bidgol endorheic lake in Iran and sequenced a selection of them. Multilocus Sequence Analysis (MLSA) and Average Nucleotide Identity (ANI) revealed multiple clusters (phylogroups) of organisms present in the lake. Patterns of intein and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) presence/absence and their sequence similarity, GC usage along with the ANI and the identities of the genes used in the MLSA revealed that two of these clusters share an exchange bias toward others in their phylogroup while showing reduced rates of exchange with other organisms in the environment. However, a third cluster, composed in part of named species from other areas of central Asia, displayed many indications of variability in exchange partners, from within the lake as well as outside the lake. We conclude that barriers to gene exchange exist between the two purely Aran-Bidgol phylogroups, and that the third cluster with members from other regions is not a single population and likely reflects an amalgamation of several populations.

Keywords: Halobacteria, Multilocus Sequence Analysis (MLSA), Average Nucleotide Identity (ANI), intein, CRISPR

INTRODUCTION

Besides an obligate requirement for high concentrations of NaCl, a unifying trait of Halobacteria (often referred to colloquially as the haloarchaea)—a class within the archaeal phylum Euryarchaeota, is their propensity for horizontal gene transfer (HGT) (Legault et al., 2006; Rhodes et al., 2011; Nelson-Sathi et al., 2012; Williams et al., 2012). Although HGT occurs continuously, events that provide an adaptive advantage and are maintained in modern lineages can be detected. For instance, HGTs from bacterial lineages into the Halobacteria occurred before their last common ancestor and brought respiration and nutrient transport genes that transformed them from a methanogen to their current aerobic heterotrophic state (Nelson-Sathi et al., 2012). Other examples including rhodopsins (Sharma et al., 2006), tRNA synthetases (Andam et al., 2012), 16S rRNA genes (Boucher et al., 2004), membrane proteins (Cuadros-Orellana et al., 2007), and genes allowing the assembly of novel pathways (Khomyakova et al., 2011) have been reported for this group and reflect the adaptive benefit of acquiring these genes.

HGT into the Halobacteria has profoundly impacted their evolution; however, understanding this contribution is only part of their evolutionary picture. The study of recombination frequency among this class has been utilized to address population genetics questions that address whether they are clonal (i.e., linked alleles at different loci) or "sexual" in the sense that alleles at different loci are randomly associated. Several studies have addressed those questions by assessing the impact of frequent HGT on Halobacteria. Homologous replacement of loci was inferred within and between phylogenetic clusters (phylogroups) using Multilocus Sequence Analysis (MLSA) on closely related strains (Papke et al., 2004) and comparative analyses of genomes (Williams et al., 2012). Within phylogroups where genetic diversity was less than one percent divergent for protein coding genes, alleles at different loci were randomly associated whereas between phylogroups they were not (Papke et al., 2007) indicating haloarchaea are highly sexual. Measurements of frequency across the breadth of halobacterial diversity indicates no absolute barrier to homologous recombination; rather between relatives, there is a log-linear decay in recombination frequency relative to phylogenetic distance (Williams et al., 2012).

Laboratory experiments also support these results. Mating experiments measuring the rate of recombination using *Haloferax (Hfx) volcanii* and *Hfx. mediterranei* auxotrophs demonstrated the degree of genetic isolation between species was much lower than expected. The observed rate of exchange between species suggested that given an opportunity over time these species would homogenize, indicating strong barriers to recombination would have to exist for speciation to occur, and for lineages to be maintained (Naor et al., 2012). Further, mating experiments demonstrated that enormous genomic fragments (i.e., 300–500 kb, ~18% of the chromosome size) could be exchanged in a single event (Naor et al., 2012). Similar large fragment exchange events were recently observed in natural isolates from Deep Lake (Antarctic hypersaline lake): Distantly related strains

(<75% average nucleotide identity) shared up to 35 kb with nearly 100% sequence identity (DeMaere et al., 2013).

The Halobacteria have clearly been shaped by gene transfer and are actively engaged in substantial genetic exchange. However, little is known about genomic diversity within populations, and the impact of gene flow is unknown at these scales. In this study we report the intra and inter population sequence diversity of *Halorubrum* spp. strains cultivated from the same location and compare them to the genomic diversity of type strains from the same genus. Our results lead to insights on the genomic diversity that comprises haloarchaeal species.

METHODS

GROWTH CONDITIONS AND DNA EXTRACTION

Halorubrum spp. cultures were grown in Hv-YPC medium (Allers et al., 2004) at 37°C with agitation. DNA from Halobacteria was isolated as described in the Halohandbook (Dyall-Smith, 2009). Briefly, stationary-phase cells were pelleted at 10,000 × *g*, supernatant was removed and the cells were lysed in distilled water. An equal volume of phenol was added, and the mixture was incubated at 65°C for 1 h prior to centrifugation to separate the phases. The aqueous phase was reserved and phenol extraction was repeated without incubation, and followed with a phenol/chloroform/iso-amyl alcohol (25:24:1) extraction. The DNA was precipitated with ethanol, washed, and re-suspended in TE (10 mM tris, pH 8.0, 1 mM EDTA).

MULTILOCUS SEQUENCE ANALYSIS (MLSA)

Five housekeeping genes were amplified using PCR. The loci were *atpB*, *ef-2*, *glnA*, *ppsA*, and *rpoB* and the primers used for each locus are listed in **Table 1**. To more efficiently sequence PCR products, an 18 bp M13 sequencing primer was added to the 5′ end of each degenerate primer (**Table 1**). Each PCR reaction was 20 μl in volume. The PCR reaction was run on a Mastercycler Ep Thermocycler (Eppendorf) using the following PCR cycle protocol: 30 s initial denaturation at 98°C, followed by 40 cycles of 30 s at 98°C, 5 s at the annealing temperature for each set of primers and 15 s at 72°C. Final elongation occurred at 72°C for 1 min. **Table 2** provides a detailed list of reagents and the PCR mixtures for each amplified locus. The PCR products were separated by gel electrophoresis with agarose (1%). Gels were stained with ethidium bromide. An exACTGene mid-range plus DNA ladder (Fisher Scientific International Inc.) was used to estimate the size of the amplicons, which were purified using Wizard SV gel and PCR cleanup system (Promega). The purified amplicons were sequenced by Genewiz Inc. using Sanger sequencing technology.

GENOME SEQUENCING

DNA purity was analyzed with a Nanodrop spectrophotometer, was quantified using a Qubit fluorometer (Invitrogen) and then prepared for sequencing using the Illumina Nextera XT sample preparation kit as described by the manufacturer. Fragmented and amplified libraries were either normalized using the normalization beads and protocol supplied with the kit, or manually as described in protocols for the Illumina Nextera kit. Libraries were loaded onto 500 cycle MiSeq reagent kits with a 5% spike-in PhiX control, and sequenced using an Illumina MiSeq benchtop sequencer. The genomes to be sequenced were selected based upon the results of the initial PCR MLSA data analysis (see Results).

Table 1 | Degenerate primers used to PCR amplify and sequence the genes for MLSA.

	MLSA primer sequence 5′–3′	
Locus	**Forward**	**Reverse**
atpB	tgt aaa acg acg gcc agt aac ggt gag scv ats aac cc	cag gaa aca gct atg act tca ggt cvg trt aca tgt a
ef-2	tgt aaa acg acg gcc agt atc cgc gct bta yaa stg g	cag gaa aca gct atg act ggt cga tgg wyt cga ahg g
glnA	tgt aaa acg acg gcc agt cag gta cgg gtt aca sga cgg	cag gaa aca gct atg acc ctc gcs ccg aar gac ctc gc
ppsA	tgt aaa acg acg gcc agt ccg cgg tar ccv agc atc gg	cag gaa aca gct atg aca tcg tca ccg acg arg gyg g
rpoB	tgt aaa acg acg gcc agt tcg aag agc cgg acg aca tgg	cag gaa aca gct atg acc ggt cag cac ctg bac cgg ncc

Table 2 | PCR conditions for each locus.

	atpB	ef-2	glnA	ppsA	rpoB
Water (μl)	11.6	8.2	11.8	7.9	11.9
5× phire reaction buffer (μl)	4.0	4.0	4.0	4.0	4.0
DMSO (μl)	0.6	0	0.4	0.6	0.6
Acetamide (25%, μl)	0	4.0	0	4.0	0
dNTP mix (10 mM, μl)	0.4	0.4	0.4	0.4	0.4
Forward primer (10 mM, μl)	1.0	1.0	1.0	1.0	1.0
Reverse primer (10 mM, μl)	1.0	1.0	1.0	1.0	1.0
Phire II DNA polymerase (μl)	0.4	0.4	0.4	0.4	0.4
Template DNA (20 ng/μl, μl)	1.0	1.0	1.0	0.7	0.7
Annealing temperature (°C)	60.0	61.0	69.6	66.0	63.7

GENOME ASSEMBLY

Type strain genomes were obtained from the NCBI ftp repository. *Halorubrum lacusprofundi* and the non-*Halorubrum* genomes (*Haloarcula marismortui* ATCC 43049 and *Har. hispanica* ATCC 33960 as well as *Haloferax volcanii* DS2 and *Hfx. mediterranei* ATCC 33500) are completed projects. The other *Halorubrum* genomes are drafts, also obtained from the NCBI ftp repository. New draft genomes were sequenced using an Illumina MiSeq platform. Assembly on strain Ga2p was carried out using the ngopt A5 pipeline(Tritt et al., 2012) while all others were assembled via the CLC Genomics Workbench 6.0.5 suite with a trim and merge workflow with scaffolding enabled.

To ensure equal gene calling across the genomes all genomes, including the 19 draft and completed *Halorubrum*, *Haloferax*, and *Haloarcula* genomes available on the NCBI ftp site as of June 2013, were reannotated using the rapid annotation using subsystem technology (RAST) server (Aziz et al., 2008). Assembled contigs were reconstructed from the RAST-generated genbank files for all genomes using the seqret application of the emboss package (Rice et al., 2000).

PHYLOGENETIC METHODOLOGY

Top scoring BLASTn hits for each MLSA target gene (*atpB*, *ef-2*, *glnA*, *ppsA*, and *rpoB*) in each genome were identified. Multiple-sequence alignments (MSAs) were generated by translating the genes to protein sequences in SeaView (Gouy et al., 2010), aligning the proteins using MUSCLE (v.3.8.31) (Edgar, 2004) and then reverting back to the nucleotide sequences. In-house scripts created a concatenated alignment of all five genes. The best model of evolution was determined by calculating the Akaike Information Criterion with correction for small sample size (AICc) in jModelTest 2.1.4 (Guindon et al., 2010; Darriba et al., 2012). The best-fitting model was GTR + Gamma estimation + Invariable site estimation. A maximum likelihood (ML) phylogeny was generated from the concatenated MSA and individual gene phylogenies from the individual gene MSAs using PhyML (v3.0_360-500M)(Guindon et al., 2010). PhyML parameters consisted of GTR model, estimated p-invar, 4 substitution rate categories, estimated gamma distribution, subtree pruning, and regrafting enabled with 100 bootstrap replicates.

PAIRWISE SEQUENCE IDENTITY CALCULATION

Calculation of pairwise identities was carried out using Clustal Omega on the EMBL-EBI webserver (http://www.ebi.ac.uk/Tools/msa/clustalo/). The alignments were uploaded and percent identity matrices calculated (Sievers et al., 2011).

INTEIN METHODOLOGY

To retrieve haloarchaeal intein sequences Position-Specific Scoring Matrices (PSSMs) were created using the collection of all inteins from InBase, the Intein database, and registry (Perler, 2002). A custom database was created with all inteins, and each intein was used as a seed to create a PSSM using the custom database. These PSSMs were then used as a seed for PSI-BLAST (Altschul et al., 1997) against each of the halobacterial genomes available from NCBI. A size exclusion step was then performed to remove false positives. Inteins were then aligned using MUSCLE (Edgar, 2004) with default parameters in the SeaView version 4.0 software package (Gouy et al., 2010). Insertions, which passed the size exclusion step but did not contain splicing domains, were filtered out and the previous steps were repeated using the resulting dataset on this study's dataset. Once the collection of haloarchaeal inteins was complete, sequences were re-aligned using SATé v2.2.2 (Liu et al., 2012) to generate a final alignment.

INTEIN PHYLOGENETIC METHODOLOGY

Intein protein sequences were retrieved using in house scripts. Each intein allele was aligned separately using MUSCLE (v.3.8.31) (Edgar, 2004). In-house scripts created a concatenated alignment from the allele alignments. ProtTest v3.4 (Darriba et al., 2011) evaluated the protein sequences for an optimal model using the AICc and returned WAG_I+G+F. A presence-absence matrix of zeros and ones was amended to each taxon's alignment data. The presence-absence data allows for grouping of taxa by sharing or lacking an allele. This complements the protein data, and allows the resolution of taxa with few inteins from those lacking them entirely or possessing many. To accommodate the two different formats of data simultaneously MrBayes v3.2.2 (Ronquist and Huelsenbeck, 2003; Ronquist et al., 2012) was employed for the phylogenetic reconstruction.

AVERAGE NUCLEOTIDE IDENTITY/TETRAMER ANALYSIS

JSpecies1.2.1 (Richter and Rosselló-Móra, 2009) was used to analyze the genomes for Average Nucleotide Identity (ANI) and tetramer frequency patterns. As the relationships of interest for this study are within the same genus only the nucmer and tetra algorithms were used. The BLAST-based ANI was not used as we were primarily interested in understanding the degree of relatedness between closely related organisms, which the nucmer method is equally capable of (Richter and Rosselló-Móra, 2009). Additionally, the increased rate of drop-off between moderately divergent sequences (<90%) the nucmer method yields relative to the BLAST method (Richter and Rosselló-Móra, 2009) was useful in highlighting when organisms were dissimilar. The default settings for both algorithms were used (Richter and Rosselló-Móra, 2009).

CODON POSITION GC CONTENT

Complete sets of nucleotide sequences for all called ORFs were downloaded from RAST. In house scripts confirmed that all ORF calls were divisible by three and thus could be taken as in-frame. In house scripts were used to calculate the GC percentages for each codon position in each genome. Two-tailed *t*-tests were calculated using the StatsPlus software package (AnalystSoft, 2009).

CRISPRs

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) presence/absence patterns were determined using the CRISPR Recognition Tool (CRT) v1.2 (Bland et al., 2007) with minimum repeat and minimum spacer parameters set to 30 nucleotides. All other parameters were the CRT defaults.

RESULTS

ASSEMBLED GENOMES

The assembled genomes ranged in size from 2.3 to 4.2 Mb. The median assembled genome size is 3.6 Mb. The median N50 (the size of the contig where 50% of the basepairs in the assembly are part of a contig that size or larger. N75 and N90 are similar but use 75 and 90% cutoffs) was 47.5 kb with a range from 1.86 to 80.3 kb (see **Table 3**, for statistics on the assembled genomes). Plasmids were not identified during assembly. As such, if some isolates possess differing numbers or types of plasmids then some of the genome-to-genome size variability may be attributable to this. A list of genomes used in this study can be found in **Table 4**.

PHYLOGENETIC ASSIGNMENT OF PHYLOGROUPS

Initial MLSA analysis (5-genes: *atpD*, *ef-2*, *glnA*, *radA*, *rpoB*) revealed the presence of three well-supported clusters [hereafter referred to as phylogroups *in sensu* (Papke et al., 2007)] within the canonical *Halorubrum* population of Aran-Bidgol (**Figures 1**, **2**). A phylogroup was initially defined as a cluster of isolates with very low sequence divergence across the sequenced (MLSA) loci (<~1%). Seventeen of these isolates were then selected for genome sequencing for a higher resolution assessment. Selection criteria were biased toward the two larger phylogroups (A and B) to facilitate comparison between clusters. Only a single genome

Table 3 | Assembly statistics for the genomes sequenced in this study.

	C191	C3	C49	Cb34	E3	E8	Ea1	Ea8	Eb13	Ec15	Fb21	G37	Ga2p	Ga36	Hd13	Ib24	LD3	LG1
N75 (kb)	18.9	2.3	23.2	24.7	1.1	13	30.0	25.1	25.4	42.7	25.3	272	41.1	23.8	32.1	23.2	21.4	8.4
N50 (kb)	54.9	4.4	56.3	42.9	1.9	2.3	43.8	5.16	51.6	80.3	42.7	68.1	74.9	51.2	64.4	43.4	39.6	32.1
N25 (kb)	97.3	7.8	99.8	73.4	3.5	4.0	77.5	95.4	95.7	131.8	90.3	118.4	118.9	91.9	83.0	68.2	76.0	67.9
Minimum (kb)	0.5	0.4	0.5	0.5	0.4	0.4	0.5	0.5	0.5	0.6	0.5	0.5	0.3	0.5	0.5	0.5	0.5	0.4
Maximum (kb)	190.2	40.5	183.6	123.4	26.7	25.0	203.3	169.6	268.1	412.4	174.7	230.0	246.3	145.6	122.0	190.3	145.8	153.4
Average (kb)	16.6	2.9	22.5	23.1	1.5	1.8	24.7	22.6	23.3	44.3	20.6	25.7	40.3	21.0	27.9	19.6	17.5	4.4
Contig count	233	1165	159	145	2764	1278	159	166	156	74	176	138	83	160	137	189	213	1090
Length (Mb)	3.87	3.33	3.58	3.35	4.21	2.26	3.93	3.75	3.63	3.28	3.63	3.55	3.35	3.36	3.82	3.70	3.73	4.79
Base composition (GC%)	66.0	65.8	65.8	67.6	65.5	66.3	67.0	67.6	67.5	67.6	66.6	67.1	67.8	67.7	67.6	67.6	66.2	66.0
Number of coding sequences	3908	3379	3529	3323	4147	2187	3977	3672	3544	3245	3600	3617	3400	3382	3718	3612	3724	4615
Number of RNAs	57	37	49	54	51	31	50	49	48	47	65	48	49	47	51	48	56	69

from phylogroup C was sequenced. Once genomic data were available, the PCR amplicons were replaced with the full-length genes from the assemblies. Further analysis made use of only these genomic sequences. The addition of the 19 NCBI genomes was made to provide context to the placement of the phylogroups within the genus and to determine their relationship with each other. The phylogenetic reconstruction including the type strains sequences revealed the presence of a fourth phylogroup (designated D) composed of three isolates from Aran-Bidgol and five type strains isolated from Central Asia and China (**Figure 2**).

PHYLOGROUPS A AND B ARE WELL-SUPPORTED AS DISCRETE AND COHESIVE ENTITIES

The bootstrap values provided by the phylogenetic reconstruction strongly supported both phylogroups A and B. Individual gene trees and the concatenated gene tree returned support values of 99% or higher for all of the clusters (**Figures 1**, **2**) and the trees showed no paraphyly with other taxa. Both phylogroups also displayed sequence divergence below 1% across the five loci (**Table 5**). Further, genome-level analysis (ANI) demonstrated similar results to the MLSA data (**Figure 3**). Additional support for these phylogroups came from the tetramer frequency analysis, which found no discordance amongst the members of either group, and each phylogroup displayed an intra-group ANI ≥98%. An analysis of G+C composition in the protein coding ORFs found that the strains within phylogroups A and B had a statistically different content in overall coding G+C and at the third codon position ($P < 0.05$ for both, **Figure 4**). Analyses of the inter-phylogroup differences showed the two phylogroups were quite different from each other and all other examined taxa. Both clusters were less than 97% similar in their pairwise MLSA distance to any other taxon in this study. Additionally phylogroups A and B were different from each other in tetramer frequency (below the 0.9900 correlation of Richter and Rosselló-Móra, 2009), ANI (only ~87% identity), and G+C content in the third codon position ($P < 0.05$; two-tailed *t*-test, **Figure 4**). Taken together these data support the notion that these phylogroups are discrete entities within a single environment, and that the individual phylogroups are cohesive.

To further evaluate the cohesion of the phylogroups a survey of inteins was performed. Inteins are molecular parasites that invade new hosts through horizontal transmission (Okuda et al., 2003; Swithers et al., 2013). Their patterns of presence and absence have been used as a barometer for horizontal transfer between closely and distantly related lineages (Swithers et al., 2013). Analysis of intein distributions supported earlier findings of cohesion within phylogroups and major distinctions between the phylogroups (**Figure 5**). Phylogroup A contains three non-fixed intein alleles that are present in more than half of the isolates, *cdc21*a, *cdc21*b, and *pol-II*a. Phylogroup B contains four non-fixed intein alleles also present in half or more of its isolates, *rir1*-b, *rfc*-a, *polB*a, and *polB*b but are absent from phylogroup A. Closer examination of the two shared alleles reveals that these inteins are not the same between the phylogroups. The *pol-II*a inteins in phylogroup B are 515aa long while those in phylogroup A are 494aa long, indicating an insertion or deletion event occurred in one of the phylogroups before the intein spread through the population. The preservation

Table 4 | List of genomes used in this study.

Organism name	NCBI identifier	Sequence source	Isolation site	Environment	Status
Haloarcula hispanica ATCC 33960	PRJNA72475	NCBI	Alicante, Spain	Solar saltern	Complete
Haloarcula marismortui ATCC 43049	PRJNA57719	NCBI	Dead Sea, Israel	Saline lake/sea	Complete
Haloferax mediterranei ATCC 33500	PRJNA167315	NCBI	Alicante, Spain	Solar saltern	Complete
Haloferax volcanii DS2	PRJNA46845	NCBI	Dead Sea, Israel	Saline lake/sea	Complete
Halorubrum sp. T3	PRJNA199598	NCBI	Yunnan, China	Solar saltern	Draft
Halorubrum aidingense JCM 13560	PRJNA188616	NCBI	Xin-Jiang, China	Saline lake	Draft
Halorubrum arcis JCM 13916	PRJNA188617	NCBI	Xin-Jiang, China	Saline lake	Draft
Halorubrum californiensis DSM 19288	PRJNA188618	NCBI	California, United States	Solar saltern	Draft
Halorubrum coriense DSM 10284	PRJNA188619	NCBI	Geelong, Australia	Solar saltern	Draft
Halorubrum distributum JCM 10118	PRJNA188621	NCBI	Turkmenistan	Saline soils	Draft
Halorubrum distributum JCM 9100	PRJNA188620	NCBI	Turkmenistan	Saline soils	Draft
Halorubrum hochstenium ATCC 700873	PRJNA188622	NCBI	California, United States	Solar saltern	Draft
Halorubrum kocurii JCM 14978	PRJNA188615	NCBI	Inner Mongolia, China	Saline lake	Draft
Halorubrum lacusprofundi ATCC 49239	PRJNA58807	NCBI	Deep Lake, Antarctica	Saline lake	Complete
Halorubrum lipolyticum DSM 21995	PRJNA188614	NCBI	Xin-Jiang, China	Saline lake	Draft
Halorubrum litoreum JCM 13561	PRJNA188613	NCBI	Fujian, China	Solar saltern	Draft
Halorubrum saccharovorum DSM 1137	PRJNA188612	NCBI	California, United States	Solar saltern	Draft
Halorubrum tebenquichense DSM 14210	PRJNA188611	NCBI	Atacama, Chile	Solar saltern	Draft
Halorubrum terrestre JCM 10247	PRJNA188610	NCBI	Turkmenistan	Saline soils	Draft
Hrr. Cb34	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. C49	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. Ea1	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. Eb13	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. Ib24	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. Ea8	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. Hd13	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. C3	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. E8	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. E3	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. LG1	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. Fb21	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. Ga2p	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. G37	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. LD3	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. Ec15	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft
Hrr. Ga36	PRJNA232799 (in submission)	This study	Aran-Bidgol, Iran	Saline lake	Draft

of the insertion or deletion within the phylogroups indicates that gene flow is occurring more readily within phylogroups than between, even when the same intein allele is shared. In accordance with earlier evidence, within phylogroups the intein sequence similarity is much higher than between phylogroups. It is unlikely that intein lengths are the result of sequencing or assembly artifacts, as they are constant within phylogroups.

The phylogenetic reconstruction derived from the combined presence-absence data and intein sequence data (**Figure 6**) shows clustering among phylogroup A and B of their constituent taxa. None of the taxa placed anywhere else but with the other members of its phylogroups and the posterior probabilities for these placements are high (0.991 for A and 0.923 for B). These results indicate that inteins are diverging mainly along cluster boundaries, as phylogroups A and B are distinct and separate, which further suggests that it is more challenging for the inteins to migrate outside compared to inside their phylogroups.

Another genetic element that serves to distinguish phylogroups A from B is the relative presence of CRISPRs. CRISPRs are a type of microbial innate immunity that provides a record of MGEs previously encountered by the lineage that carries them. This record serves the organism by recognizing and destroying sequences that resemble previously encountered MGEs. CRISPRs have been reported in 90% of surveyed archaeal genomes (Kunin et al., 2007), thus the presence and similarity of CRISPR loci provides a means for comparing the phylogroups. The distribution of CRISPRs was surprisingly patchy in phylogroup A and the genus as a whole; however, even more surprisingly was that putative CRIPSRs were absent in phylogroup B indicating its members may be devoid of them entirely (**Figure 5**). To assess if the absence

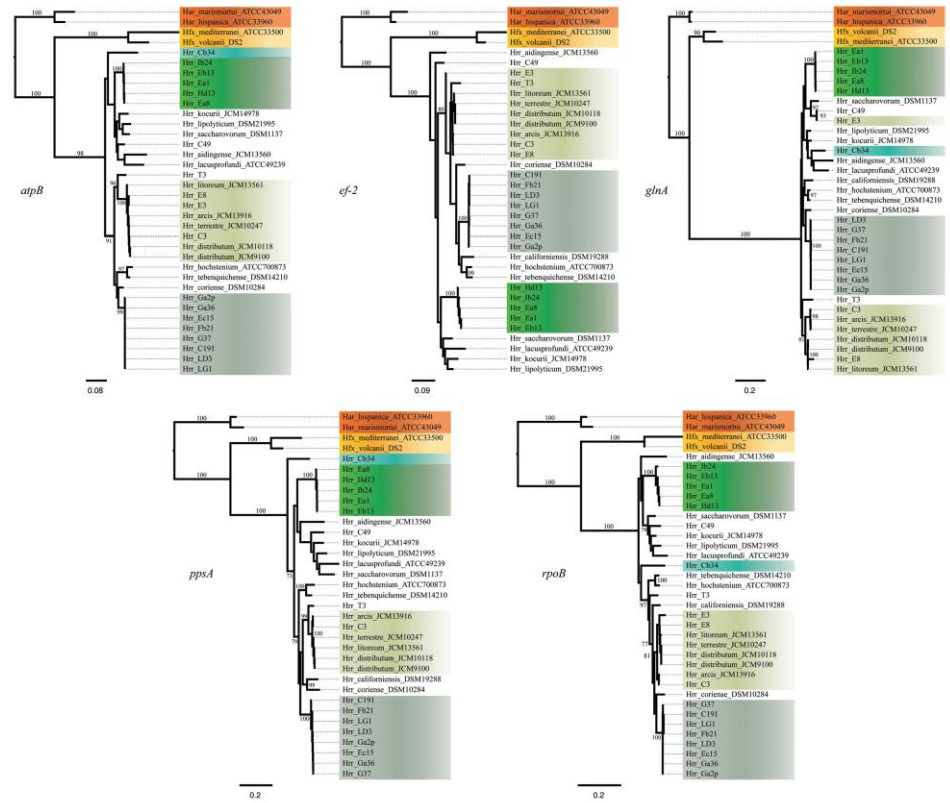


FIGURE 1 | Maximum-likelihood gene trees made from the DNA sequences of *atpB, ef-2, glnA, ppsA,* and *rpoB*. Support values on branches are bootstrap replicates. Bootstrap values below 70 are not displayed.

of CRISPRs was an artifact of using draft genome assemblies, we tested for a correlation by relating N50 to CRISPR counts per genome and found there to be no correlation ($R^2 = 0.105$, $P > 0.05$). Therefore, the CRISPR absences do not appear to be a result of genome assembly.

PHYLOGROUP D IS NOT A COHESIVE AND DISCREET ENTITY

Phylogroup D appeared in the phylogenetic reconstructions of MLSA genes after the inclusion of the NCBI *Halorubrum* genomes. It includes five genomes representing four previously described *Halorubrum* species (*Hrr. arcis*, *Hrr. terrestre*, *Hrr. Distributum,* and *Hrr. litoreum*). It was surprising that multiple named species formed such a unit, but evidence suggests it is not discreet and cohesive like phylogroups A and B: much of the data conflict leading to an ambiguous demarcation of its boundary (see below).

The phylogenetic reconstruction of this cluster is supported by the bootstrap values, with exceptions. The concatenated phylogeny has a bootstrap value of 100 at its base and the individual gene trees each support the cluster with bootstrap value of greater than 80 (**Figures 1**, **2**). Pairwise identity between the MLSA genes shows phylogroup D meets the initial criterion of <1% sequence divergence (**Table 5**). While high, the intra-cluster sequence identity is statistically lower than both phylogroup A and B values ($P < 0.05$, two-tailed *t*-test). ANI gives similar results to the pairwise identity (**Figure 3**): the intra-cluster value is ~97%. However some members of the group do not meet the 96% threshold identity, such as E3. Tetramer analysis shows good cohesion within the group, as all but one genome (E3) passed the cutoff. Both E3 and *Hrr. litoreum*'s tetramer frequency patterns are poorly correlated and are below the 0.99 coefficient cutoff advocated by the JSpecies 1.2.1 (Richter and Rosselló-Móra, 2009) package.

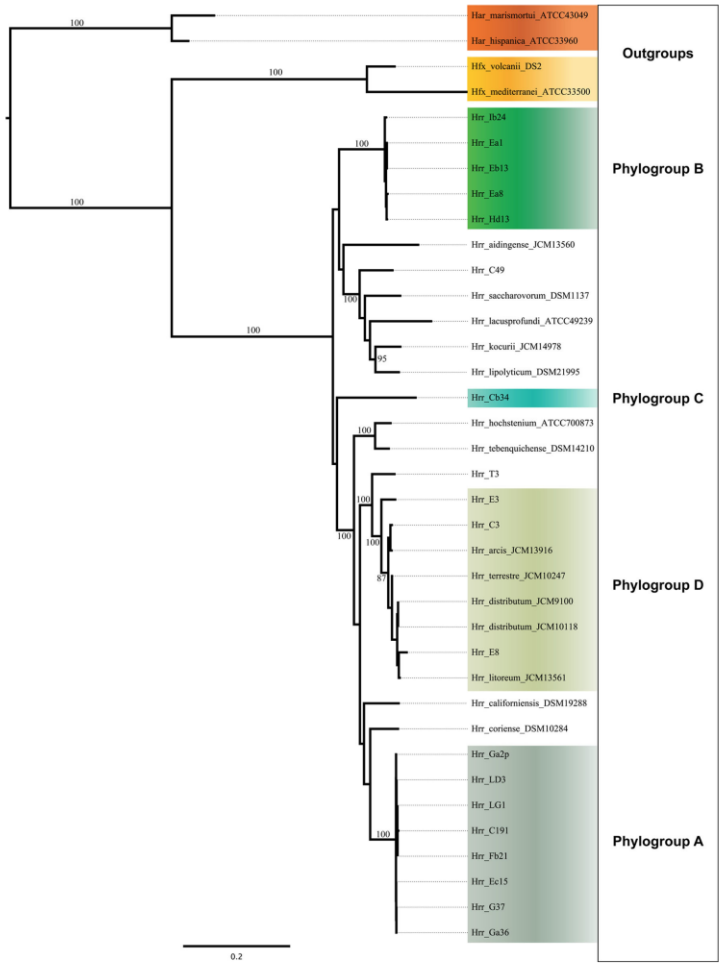
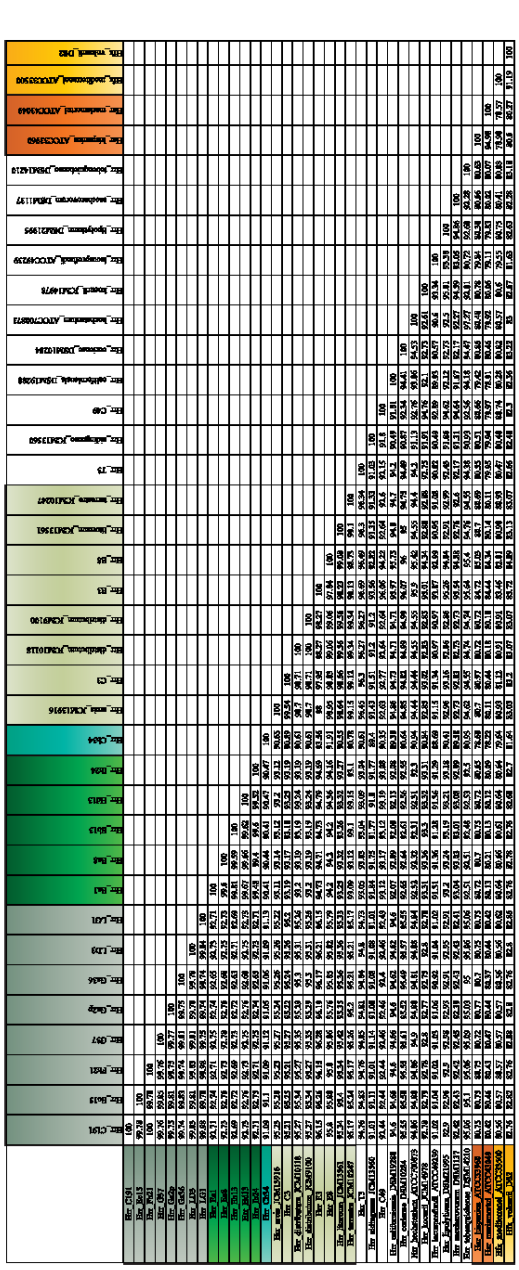


FIGURE 2 | Maximum-likelihood tree made from the concatenated DNA sequences of five housekeeping genes (*atpB, ef-2, glnA, ppsA,* and *rpoB*). Support values on branches are bootstrap replicates. Bootstraps values below 70 are not displayed.

As tetramer patterning is largely a granular filter, it strongly suggests that E3 and *Hrr. litoreum* may be distantly related, which is further supported by the ANI analysis.

The phylogroup D intein distribution patterns and sequences identities are dissimilar to phylogroup A and B (**Figure 5**). The intra-phylogroup identity of *pol-II*a is quite low in D compared to phylogroups A and B (∼78 vs. ∼99% and ∼89%, respectively). The inter-group identities are much higher between B and D than in any other phylogroup relationship (∼71%). These relationships are partly explained by *Hrr. terrestre*, which features an intein of much greater length and sequence divergence than the other alleles. This intein shares no more than 55% identity with any other phylogroup D *pol-II*a allele. If it is removed from consideration, the phylogroup D intra-cluster identity increases to ∼99%. The relatedness to phylogroup A rises to ∼53% while the value to phylogroup B is 76%. Intra-phylogroup D *cdc21b*

Table 5 | Pairwise distances of the concatenated alignment of the five MLSA genes.



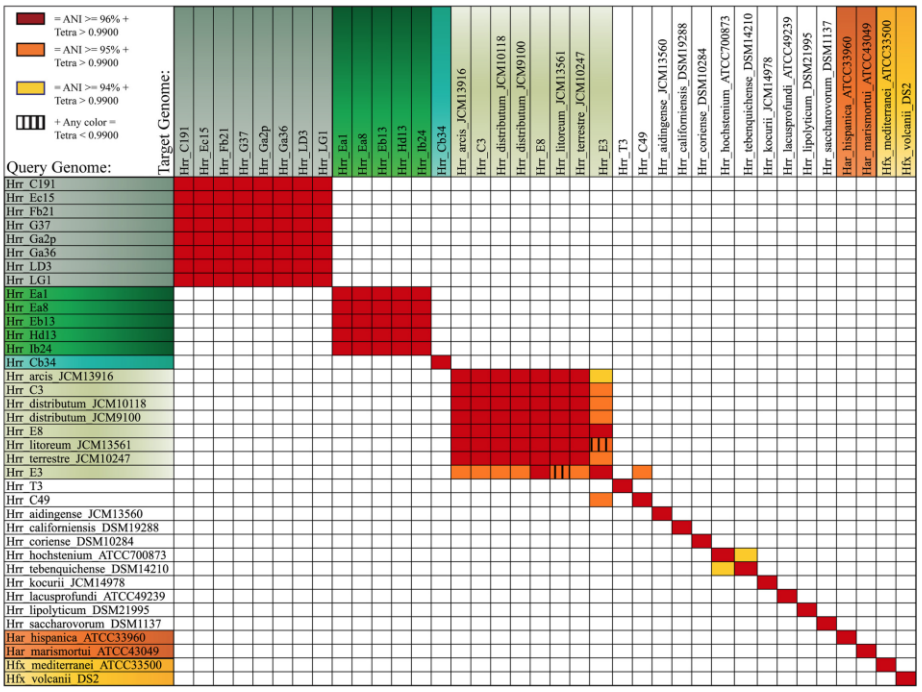


FIGURE 3 | Average Nucleotide Identity (ANI) and tetramer frequency correlation analysis. Color coding reflects three described ANI cutoffs for species delineation. Red squares represent ANI values of 96% or greater, Orange 95% or greater, and yellow represents 94% or greater. The vertical stripes indicate tetramer regression coefficients lower than 0.9900.

FIGURE 4 | GC usage of all annotated ORFs within and between phylogroups.

diversity is nearly the same as its inter-phylogroup D diversity, which further indicates phylogroup D is a fuzzy entity. The intra-phylogroup identity for the *cdc21b* intein is ∼91% (as compared to ∼100% for A and ∼99% for B) and its inter-phylogroup values are not much lower with D vs. B at ∼83% and D vs. A at ∼87%. However, the remaining taxa (*Hrr. arcis*, *Hrr. litoreum*, *Hrr. distributum*, *Hrr. terrrestre*, E8, and C3), including the named species appear to form a stable phylogroup. These data suggest that phylogroup D as constructed in our analysis is an amalgamation of populations that resembles other analyzed phylogroups but is not a cohesive unit upon additional investigation. The phylogenetic reconstruction derived from the combined presence-absence data and intein sequence data (**Figure 6**) shows that phylogroup D does not retain monophyly. Members place at four locations in the tree. The phylogroup displays high identities for core members, but "fringe" members are at the edge of inclusion.

Hrr. T3 and E3 presented significant challenges to defining the boundary of phylogroup D. As mentioned above, *Hrr.* T3 placed directly sister to the phylogroup in three of five gene phylogenies and inside the group in a fourth (**Figure 1**). In the fifth phylogeny it placed several nodes away from the cluster. The concatenation also places it sister to the cluster with maximum bootstrap support. However, its branch is long relative to the phylogroup. As noted, the pairwise identities and ANI values (**Figure 3**) both

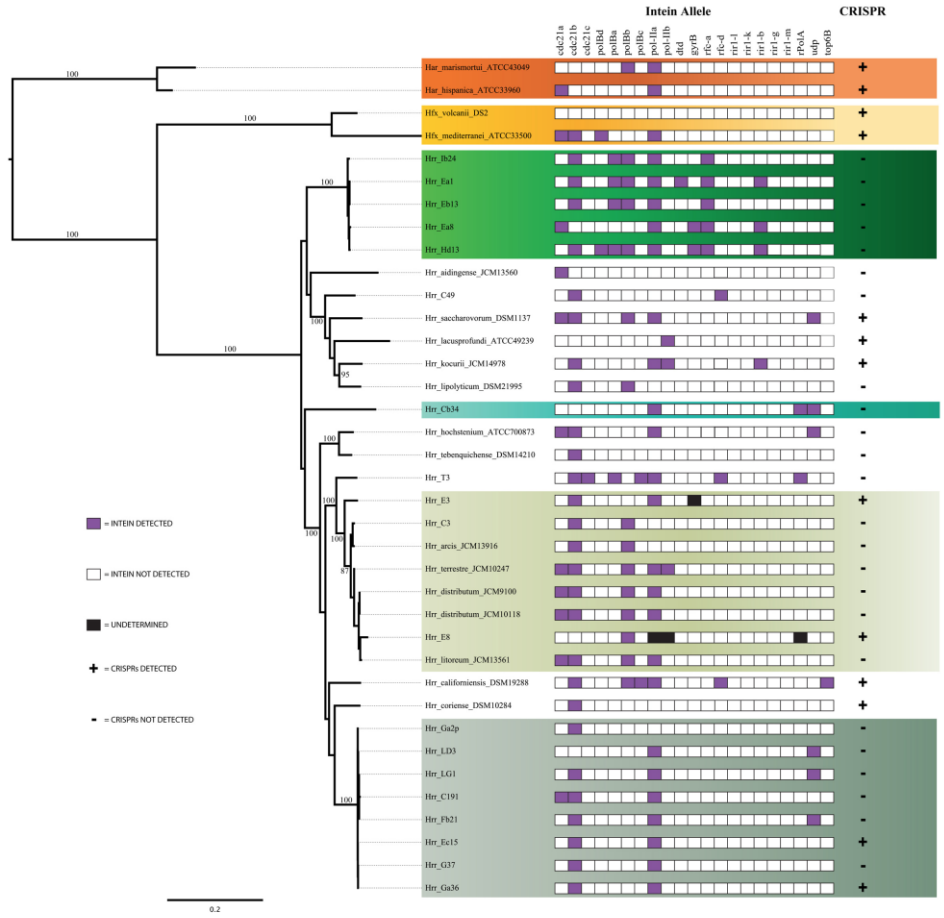


FIGURE 5 | Assessment of the presence of inteins and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs). For inteins, purple boxes indicates the presence of an intein allele, white indicates its absence and black indicates an undetermined result. For CRISPRs a (+) indicates the presence of one or more CRISPRs and a (–) indicates the absence of CRISPRs.

place it below the values seen inside the cluster. These notably lower values were used to exclude this taxon from the phylogroup. *Hrr.* E3 is less of a clean-cut case. Its *glnA* gene is outside of the phylogroup. It also falls on a branch by itself at the base of the cluster with rest of the phylogroup supported by an 87% bootstrap score. However, its intra-cluster pairwise and ANI values are several percent higher than *Hrr.* T3 and only a percent or two below most of the other members of the phylogroup. Overall, the ANI support was on the edge of current cutoffs for species delineation (95% or 96%) (Konstantinidis et al., 2006; Richter and Rosselló-Móra, 2009). Its genome had ANIs ∼95% to most of the others in the phylogroup and was only 94% to *Hrr. arcis*. Further, E3's tetramer frequency was also substantially different from *Hrr. litoreum*. A possible explanation for some of these differences is that C49 and E3 show a high degree of sequence identity (95% ANI). It is also C49 with which E3's *glnA* gene associates. Finally, the combined presence-absence and intein phylogeny places these taxa together (**Figure 6**). These data suggest that the two lineages may have engaged in a recent round of genetic exchange, which might explain why E3 is on the periphery

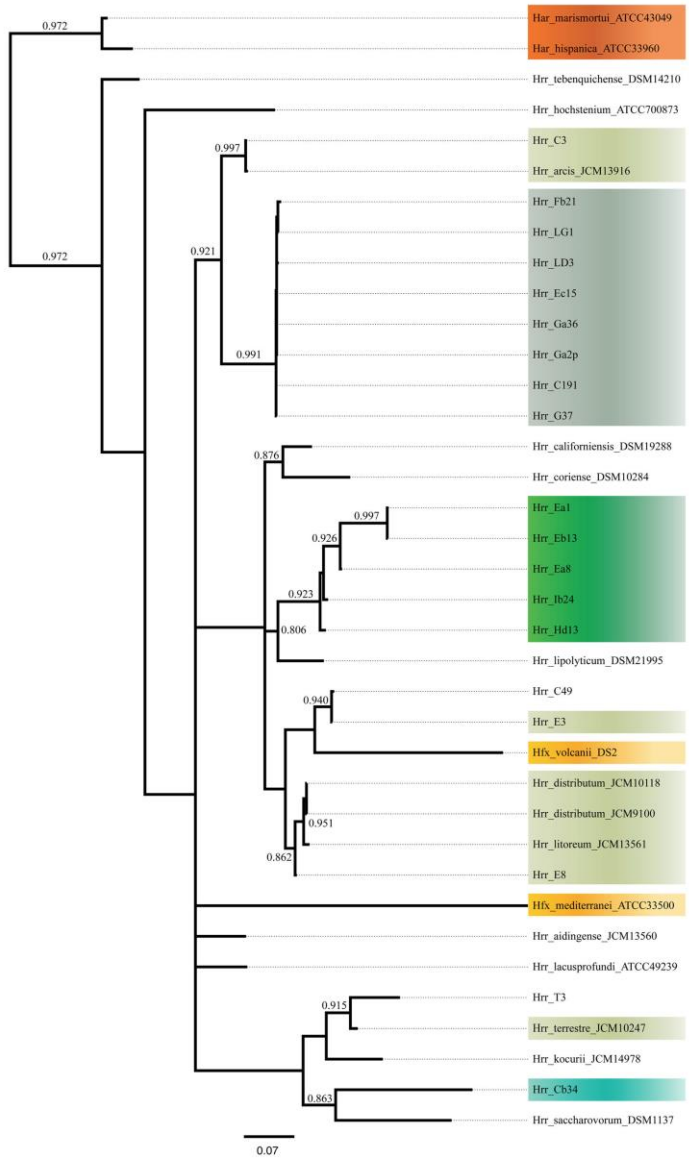


FIGURE 6 | Bayesian tree made from presence-absence of intein alleles and protein sequences of present alleles. Support values on branches are posterior probabilities. Posteriors below 0.8 are not displayed.

of phylogroup D. Ultimately, it was concluded to include E3 as a member of the phylogroup with the acceptance that it was probably an arbitrary distinction in either direction. It was this difficulty in defining the border that resulted in closer examination of phylogroup D and the ultimate rejection of it representing the same sort of entity that phylogroups A and B are.

DISCUSSION

ARE PHYLOGROUPS SPECIES?

The data presented here raise the question: are phylogroups species? We use the term "phylogroup" because a polyphasic analysis (currently defined for the Halobacteria by Oren and Ventosa, 2013) for species description has yet been published on any of the clusters. Still, an evaluation of the data strongly suggests that at least some phylogroups will be eventually described as new species. From the phylogenetic data the perspective provided by the type strain sequences would indicate that phylogroups A and B are unique species. The ANI data support the idea of phylogroups A and B belonging to separate, novel species as several studies advocate cutoffs for species delineation (Konstantinidis and Tiedje, 2005; Konstantinidis et al., 2006; Richter and Rosselló-Móra, 2009) and phylogroups A and B meet all of them. Additionally, both phylogroups form a cohesive cluster with no particular affinity for other clusters, as evidenced by the strong bootstrap support at the base of each cluster. Also, phylogroups A and B are separated from the others by multiple type strains that place between them. Despite many of these branches being poorly supported, their placement and the strong cohesion within the phylogroups argue that the clusters indicate meaningful phylogenetic splits. These splits likely represent barriers that affect the frequency of gene flow between phylogroups, but not within.

Despite the phylogroups' seemingly species-like attributes, each gene analyzed demonstrates a different topological relationship for them, which means species cannot be viewed as a group of individuals that have a common ancestor, as would be expected from eukaryotic species. While the individual organisms in a prokaryotic species do not share a common ancestor, some of their genes will. For instance, analysis of marine *Vibrio* strains showed that ~1% of the genes within populations shared a common heritage (Shapiro et al., 2012), thus the term species in prokaryotes reflects a process of homogenization, but not heritage, the assumption of Darwinian tree-like speciation. A model that could explain the data is that genes are recombined frequently within *Halorubrum* populations and less so between them. Within the high frequency recombination background new genes that confer selective advantage constantly enter phylogroups from outside the population. These advantageous genes/alleles rise rapidly in frequency throughout the recombining population causing them to diverge in comparison to other phylogroups, yet remaining homogenized within. Like continental drift gives the appearance of discreet units yet are comprised of parts derived from other continents, so too are these two *Halorubrum* phylogroups.

Phylogroup D demonstrates further the model above, as recombination from outside the group is causing divergence, and disallowing a clean species prediction compared to phylogroups A or B. Therefore, phylogroups D is unlikely to be a single species because it is less cohesive in other measurements, which reflects that it contains several previously described species and also that it has engaged in numerous gene exchanges with not-to-distantly-related organisms. Alternatively, since species assignment is a pragmatic endeavor it could be argued from our data and analyses that phylogroup D is a single species with more genetic diversity than found in A and B. The ambiguous relationships of *Hrr.* T3 and E3 suggest there are different recombination partners available to the cluster members. Such differential exchange partners are key elements in microbial speciation (Papke and Gogarten, 2012) and it could be that T3 and E3 are in the process of speciation from the other members of D, but is incomplete. Tetramer frequency data, which has been demonstrated to convey phylogenetic information (Bohlin et al., 2008a,b) casts doubt on the phylogroup representing a single species. It is less stringent than ANI, being more inclusive with the clusters it forms at typical cutoff values (Richter and Rosselló-Móra, 2009). For this reason, when tetramer frequencies are in disagreement it is likely that the two sequences being compared are not closely related. Thus, the tetramer frequency difference between E3 and *Hrr. litoreum* is also strong evidence for those two taxa not belonging to the same species. Interestingly, if T3 and E3 belong to different species and are removed from consideration, the remaining members of phylogroup D would be a single species by all measurements and cutoffs, and yet are still comprised of four named species. However, these strains were isolated from three different geographic regions of Asia at three different time points (Zvyagintseva and Tarasov, 1987; Ventosa et al., 2004; Cui et al., 2007; Xu et al., 2007), from Chinese solar salterns to Turkmenistani saline soils. While the role of geography and ecology in haloarchaeal speciation is unsettled (Oh et al., 2010; DeMaere et al., 2013; Dillon et al., 2013; Zhaxybayeva et al., 2013) all four of the named species have undergone polyphasic characterization, including DNA-DNA hybridization (Ventosa et al., 2004; Cui et al., 2007; Xu et al., 2007). Presumably, if these taxa lived in the same environments and exchanged genes with each other in a positively biased manner like phylogroups A and B, they would be homogenized and indistinguishable by current polyphasic description processes. What sets phylogroup D apart in our analysis is that we do not have population data on members from the same site, and cannot compare equivalently: if we had more data from natural populations like we do for phylogroups A and B, it might be possible to detect reliable differences that separate the named species into different MLSA phylogroups. For example, dozens of *Sulfolobus* strains isolated from geographically distant sites were less than 1% divergent across multiple loci, yet population data analysis demonstrated they fall into discreet clusters associated with geography (Whitaker et al., 2003) While the taxonomy of the Halobacteria is in flux (for example: McGenity and Grant, 1995; Oren and Ventosa, 1996) it seems unlikely that these four separate species will be merged into one. Recent work has served to split *Hrr. terrestre* from *Hrr. distributum* (Ventosa et al., 2004). Thus, it is challenging to conceive of phylogroup D as a single species, which serves as a strong example of the limits

to MLSA and ANI in regards to being the defining measurements of species.

CRISPR DISTRIBUTION MAY BE THE RESULT OF SELECTION

It is important to acknowledge that the patchy CRISPR distribution may be in part an artifact of genome assembly. Repeats can prove a challenge to assembly of short read data (Miller et al., 2010; Magoc et al., 2013) and CRISPRs are repeat heavy. However, false negatives that may exist are unlikely to be directly correlated with assembly quality, and no significant correlation is found between N50 score and the number of CRISPR arrays detected ($P > 0.05$). Additionally, the use of a different CRISPR detector, Crass v0.3.6 (Skennerton et al., 2013), which analyzes raw sequencing reads, rather than finding them in assemblies, supported the CRISPRs reported and found only slight evidence for three additional taxa possessing CRISPRs (data not shown). This would only represent individual CRISPR repeats no larger than about three spacers. While CRISPRs this size have been reported (Kunin et al., 2007) the evidence is inconclusive and if these three taxa do possess CRISPRs their distribution would remain sparse. Only seven of the 18 genomes sequenced in this study would possess them.

CRISPRs have been reported to be very common in the archaea (Jansen et al., 2002; Godde and Bickerton, 2006; Kunin et al., 2007; Held et al., 2010) with reported incidence as high as 90% (Koonin and Makarova, 2009). The incidence in bacteria is closer to 50%. The higher incidence in the archaea may be due to the underrepresentation of archaeal genomes in databases. With viruses and other MGEs so common (for discussion of haloviruses see Dyall-Smith et al., 2003; Porter et al., 2007) and horizontal transfer of CRISPRs a frequent occurrence (Kunin et al., 2007; Sorek et al., 2008), why does selection ever conjure a no-CRISPR lineage? One possibility is that the benefit provided is not strong enough to outweigh the costs, as CRISPR systems require precise matches with their target, and a "proto-spacer" with one or two mismatches can eliminate functionality (Deveau et al., 2008). The loss of cassettes in CRISPR arrays is not uncommon (Deveau et al., 2008; Díez-Villaseñor et al., 2010; Touchon and Rocha, 2010), while loss of an entire array is less so (Held et al., 2010; Touchon and Rocha, 2010). Possession of large CRISPR arrays may not offer extra protection against the viruses in an environment (Díez-Villaseñor et al., 2010). It might be that if predation level by MGEs rise and fall then the value of the CRISPR system might follow those trends. *Escherichia* and *Salmonella* CRISPR arrays do not appear to deteriorate rapidly enough to be lost entirely and they show a high rate of transfer and loss of the *cas* proteins that form the machinery of the functional system (Touchon and Rocha, 2010). This might suggest that the need for the system may not be constant. Another reason for degradation of the system could be related to it behaving in an auto-immune fashion. When challenged by artificial constructs including a proto-spacer and a gene complementing an autotrophic defect in the strain, *Sulfolobus* cells developed a surprisingly large number of deletion mutants in the spacer providing immunity to the construct (Gudbergsdottir et al., 2011). The authors speculated that there might be some small degree of feedback where the system attacks the host's spacer in addition to that of the MGE. The cellular repair systems may then easily delete the spacer during the repair process. Feedback against self and similar to self DNA, such as targeting closely related housekeeping genes (Gophna and Brodt, 2012) could also impact mating proficiency if the CRISPR system degrades the DNA of exchange partners before it can experience recombination events. It is also important to consider that mechanisms other than CRISPRs have major roles in developing resistance to MGEs (Wilson and Murray, 1991; Bickle and Krüger, 1993; Díez-Villaseñor et al., 2010). For instance, there could be a balance between CRISPRs and restriction/modification systems where one system is lost and another replaces, or complements it such that any one anti-MGE mechanism at any moment in time is in flux.

THE ABSENCE OF INTEINS SUGGESTS BARRIERS TO RECOMBINATION BETWEEN PHYLOGROUPS

Inteins are found pervasively among the archaea (Perler, 2002). They insert into genes and once translated their splicing domains use an auto-catalytic mechanism to self-excise from the protein and re-join the two halves of the polypeptide to generate a functional protein. Inteins associate with homing endonucleases (HEN), found between the splicing domains, to allow their transmission into new hosts. HENs target highly conserved sites in highly conserved genes (Swithers et al., 2009). These HENs appear to be extremely specific in their target sequences as inteins are only found inserted among the most conserved residues of highly conserved protein coding genes (Swithers et al., 2009). Their means of dissemination from host to host is, as yet, unknown although it is clear that it relies on established methods of gene flow within a population (Goddard and Burt, 1999; Gogarten and Hilario, 2006). This suggests that if two hosts have no method of transmitting genes between themselves then the resident inteins will not cross hosts, either. Thus, the patchy distribution of inteins can be interpreted as evidence for a barrier to transfer. This is particularly relevant for the alleles that are not shared between phylogroups A and B. The presence of multiple alleles not seen in the other group argues that the allele has been unable to spread. This is not implying that members of phylogroups A and B do not exchange genes, rather, the sequence divergence and lack of intein spread implies that the recombination process is hindered relative to within group genetic exchange. Indeed, if the mating observed between different *Haloferax* species (see Naor et al., 2012) is possible then almost any sequence divergence between *Halorubrum* phylogroups is akin to a speed bump rather than a mountain in slowing the rate of genetic exchange. Additionally, studies of homologous recombination have found transfers across class-level phylogenetic distance, only at increasingly lower rates as the genetic distance increases (Vulić et al., 1997; Williams et al., 2012).

AUTHOR CONTRIBUTIONS

Matthew S. Fullmer, J. Peter Gogarten, Antonio Ventosa, and R. Thane Papke participated in the design of this study and helped to draft the manuscript. Shannon M. Soucy generated the intein data and performed the majority of the intein analysis and helped to draft the manuscript. Kristen S. Swithers performed the CRT analysis and helped to draft the manuscript. Andrea M.

Makkay and Ryan Wheeler performed the MLSA PCR. Andrea M. Makkay performed the genome sequencing. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Mohammad A. Amoozegar (University of Tehran, Iran) for allowing us to analyze the Aran-Bidgol strains, and the UConn Bioinformatics Facility for providing computing resources. This research was supported by the National Science Foundation (award numbers, DEB0919290 and DEB0830024) and NASA Astrobiology: Exobiology and Evolutionary Biology Program Element (Grant Number NNX12AD70G).

REFERENCES

Allers, T., Ngo, H.-P., Mevarech, M., and Lloyd, R. G. (2004). Development of additional selectable markers for the halophilic archaeon *Haloferax volcanii* based on the *leuB* and *trpA* genes. *Appl. Environ. Microbiol.* 70, 943–953. doi: 10.1128/AEM.70.2.943-953.2004

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

AnalystSoft. (2009). *Statistical Analysis Program for Mac OS*. Alexandria, VA: AnalystSoft Inc.

Andam, C. P., Harlow, T. J., Papke, R. T., and Gogarten, J. P. (2012). Ancient origin of the divergent forms of leucyl-tRNA synthetases in the Halobacteriales. *BMC Evolutionary Biology* 12:85. doi: 10.1186/1471-2148-12-85

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75

Bickle, T. A., and Krüger, D. H. (1993). Biology of DNA restriction. *Microbiol. Rev.* 57, 434–450.

Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C., et al. (2007). CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* 8:209. doi: 10.1186/1471-2105-8-209

Bohlin, J., Skjerve, E., and Ussery, D. W. (2008a). Investigations of oligonucleotide usage variance within and between prokaryotes. *PLoS Comput. Biol.* 4:e1000057. doi: 10.1371/journal.pcbi.1000057

Bohlin, J., Skjerve, E., and Ussery, D. W. (2008b). Reliability and applications of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes. *BMC Genomics* 9:104. doi: 10.1186/1471-2164-9-104

Boucher, Y., Douady, C. J., Sharma, A. K., Kamekura, M., and Doolittle, W. F. (2004). Intragenomic heterogeneity and intergenomic recombination among haloarchaeal rRNA genes. *J. Bacteriol.* 186, 3980–3990. doi: 10.1128/JB.186.12.3980-3990.2004

Cuadros-Orellana, S., Martin-Cuadrado, A.-B., Legault, B., D'Auria, G., Zhaxybayeva, O., Papke, R. T., et al. (2007). Genomic plasticity in prokaryotes: the case of the square haloarchaeon. *ISME J.* 1, 235–245. doi: 10.1038/ismej.2007.35

Cui, H.-L., Lin, Z.-Y., Dong, Y., Zhou, P.-J., and Liu, S.-J. (2007). *Halorubrum litoreum* sp. nov., an extremely halophilic archaeon from a solar saltern. *Int. J. Syst. Evol. Microbiol.* 57, 2204–2206. doi: 10.1099/ijs.0.65268-0

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27, 1164–1165. doi: 10.1093/bioinformatics/btr088

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* 9, 772–772. doi: 10.1038/nmeth.2109

DeMaere, M. Z., Williams, T. J., Allen, M. A., Brown, M. V., Gibson, J. A. E., Rich, J., et al. (2013). High level of intergenera gene exchange shapes the evolution of haloarchaea in an isolated Antarctic lake. *Proc. Natl. Acad. Sci. U.S.A.* 110, 16939–16944. doi: 10.1073/pnas.1307090110

Deveau, H., Barrangou, R., Garneau, J. E., Labonté, J., Fremaux, C., Boyaval, P., et al. (2008). Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.* 190, 1390–1400. doi: 10.1128/JB.01412-07

Díez-Villaseñor, C., Almendros, C., García-Martínez, J., and Mojica, F. J. M. (2010). Diversity of CRISPR loci in *Escherichia coli*. *Microbiology* 156, 1351–1361. doi: 10.1099/mic.0.036046-0

Dillon, J. G., Carlin, M., Gutierrez, A., Nguyen, V., and McLain, N. (2013). Patterns of microbial diversity along a salinity gradient in the Guerrero Negro solar saltern, Baja CA Sur, Mexico. *Front. Microbiol.* 4:399. doi: 10.3389/fmicb.2013.00399

Dyall-Smith, M. (2009). *The Halohandbook - Protocols for Haloarchaeal Genetics*. Available online at: http://www.haloarchaea.com/resources/halohandbook/index.html

Dyall-Smith, M., Tang, S.-L., and Bath, C. (2003). Haloarchaeal viruses: how diverse are they? *Res. Microbiol.* 154, 309–313. doi: 10.1016/S0923-2508(03)00076-7

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Goddard, M. R., and Burt, A. (1999). Recurrent invasion and extinction of a selfish gene. *Proc. Natl. Acad. Sci. U.S.A.* 96, 13880–13885. doi: 10.1073/pnas.96.24.13880

Godde, J. S., and Bickerton, A. (2006). The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J. Mol. Evol.* 62, 718–729. doi: 10.1007/s00239-005-0223-z

Gogarten, J. P., and Hilario, E. (2006). Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements. *BMC Evolutionary Biology* 6:94. doi: 10.1186/1471-2148-6-94

Gophna, U., and Brodt, A. (2012). CRISPR/Cas systems in archaea. *Mob. Genet. Elements* 2, 63–64. doi: 10.4161/mge.19907

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView Version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* 27, 221–224. doi: 10.1093/molbev/msp259

Gudbergsdottir, S., Deng, L., Chen, Z., Jensen, J. V. K., Jensen, L. R., She, Q., et al. (2011). Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. *Mol. Microbiol.* 79, 35–49. doi: 10.1111/j.1365-2958.2010.07452.x

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Held, N. L., Herrera, A., Cadillo-Quiroz, H., and Whitaker, R. J. (2010). CRISPR associated diversity within a population of *Sulfolobus islandicus*. *PLoS ONE* 5:e12988. doi: 10.1371/journal.pone.0012988

Jansen, R., van Embden, J. D. A., Gaastra, W., and Schouls, L. M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* 43, 1565–1575. doi: 10.1046/j.1365-2958.2002.02839.x

Khomyakova, M., Bükmez, Ö., Thomas, L. K., Erb, T. J., and Berg, I. A. (2011). A Methylaspartate cycle in haloarchaea. *Science* 331, 334–337. doi: 10.1126/science.1196544

Konstantinidis, K. T., Ramette, A., and Tiedje, J. M. (2006). The bacterial species definition in the genomic era. *Philos. Trans. R. Soc. B Biol. Sci.* 361, 1929–1940. doi: 10.1098/rstb.2006.1920

Konstantinidis, K. T., and Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 102, 2567–2572. doi: 10.1073/pnas.0409727102

Koonin, E. V., and Makarova, K. S. (2009). CRISPR-Cas: an adaptive immunity system in prokaryotes. *F1000 Biol. Rep.* 1:95. doi: 10.3410/B1-95

Kunin, V., Sorek, R., and Hugenholtz, P. (2007). Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol.* 8:R61. doi: 10.1186/gb-2007-8-4-r61

Legault, B. A., Lopez-Lopez, A., Alba-Casado, J. C., Doolittle, W. F., Bolhuis, H., Rodriguez-Valera, F., et al. (2006). Environmental genomics of "*Haloquadratum walsbyi*" in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. *BMC Genomics* 7:171. doi: 10.1186/1471-2164-7-171

Liu, K., Warnow, T. J., Holder, M. T., Nelesen, S. M., Yu, J., Stamatakis, A. P., et al. (2012). SATé-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Syst. Biol.* 61, 90–106. doi: 10.1093/sysbio/syr095

Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., et al. (2013). GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics* 29, 1718–1725. doi: 10.1093/bioinformatics/btt273

McGenity, T. J., and Grant, W. D. (1995). Transfer of *Halobacterium saccharovorum*, *Halobacterium sodomense*, *Halobacterium trapanicum* NRC 34021 and *Halobacterium lacusprofundi* to the Genus *Halorubrum* gen. nov., as *Halorubrum saccharovorum* comb. nov., *Halorubrum sodomense* comb. nov., *Halorubrum trapanicum* comb. nov., and *Halorubrum lacusprofundi* comb. nov. *Syst. Appl. Microbiol.* 18, 237–243. doi: 10.1016/S0723-2020(11)80394-2

Miller, J. R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics* 95, 315–327. doi: 10.1016/j.ygeno.2010.03.001

Naor, A., Lapierre, P., Mevarech, M., Papke, R. T., and Gophna, U. (2012). Low species barriers in halophilic Archaea and the formation of recombinant hybrids. *Curr. Biol.* 22, 1444–1448. doi: 10.1016/j.cub.2012.05.056

Nelson-Sathi, S., Dagan, T., Landan, G., Janssen, A., Steel, M., McInerney, J. O., et al. (2012). Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc. Natl. Acad. Sci. U.S.A.* 109, 20537–20542. doi: 10.1073/pnas.1209119109

Oh, D., Porter, K., Russ, B., Burns, D., and Dyall-Smith, M. (2010). Diversity of *Haloquadratum* and other haloarchaea in three, geographically distant, Australian saltern crystallizer ponds. *Extremophiles* 14, 161–169. doi: 10.1007/s00792-009-0295-6

Okuda, Y., Sasaki, D., Nogami, S., Kaneko, Y., Ohya, Y., and Anraku, Y. (2003). Occurrence, horizontal transfer and degeneration of VDE intein family in Saccharomycete yeasts. *Yeast* 20, 563–573. doi: 10.1002/yea.984

Oren, A., and Ventosa, A. (1996). A proposal for the transfer of *Halorubrobacterium distributum* and *Halorubrobacterium coriense* to the genus Halorubrum as *Halorubrum distributum* comb. nov. and *Halorubrum coriense* comb. nov., respectively. *Int. J. Syst. Bacteriol.* 46, 1180–1180. doi: 10.1099/00207713-46-4-1180

Oren, A., and Ventosa, A. (2013). Subcommittee on the taxonomy of Halobacteriaceae and Subcommittee on the taxonomy of Halomonadaceae: minutes of the joint open meeting, 24 June 2013, Storrs, Connecticut, USA. *Int. J. Syst. Evol. Microbiol.* 63, 3540–3544. doi: 10.1099/ijs.0.055988-0

Papke, R. T., and Gogarten, J. P. (2012). How bacterial lineages emerge. *Science* 336, 45–46. doi: 10.1126/science.1219241

Papke, R. T., Koenig, J. E., Rodriguez-Valera, F., and Doolittle, W. F. (2004). Frequent recombination in a saltern population of *Halorubrum*. *Science* 306, 1928–1929. doi: 10.1126/science.1103289

Papke, R. T., Zhaxybayeva, O., Feil, E. J., Sommerfeld, K., Muise, D., and Doolittle, W. F. (2007). Searching for species in haloarchaea. *Proc. Natl. Acad. Sci. U.S.A.* 104, 14092–14097. doi: 10.1073/pnas.0706358104

Perler, F. B. (2002). InBase: the Intein Database. *Nucleic Acids Res.* 30, 383–384. doi: 10.1093/nar/30.1.383

Porter, K., Russ, B. E., and Dyall-Smith, M. L. (2007). Virus–host interactions in salt lakes. *Curr. Opin. Microbiol.* 10, 418–424. doi: 10.1016/j.mib.2007.05.017

Rhodes, M. E., Spear, J. R., Oren, A., and House, C. H. (2011). Differences in lateral gene transfer in hypersaline versus thermal environments. *BMC Evolutionary Biology* 11:199. doi: 10.1186/1471-2148-11-199

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the european molecular biology open software suite. *Trends Genet.* 16, 276–277. doi: 10.1016/S0168-9525(00)02024-2

Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U.S.A.* 106, 19126–19131. doi: 10.1073/pnas.0906412106

Ronquist, F., and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574. doi: 10.1093/bioinformatics/btg180

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., Szabó, G., et al. (2012). Population genomics of early events in the ecological differentiation of Bacteria. *Science* 336, 48–51. doi: 10.1126/science.1218198

Sharma, A. K., Spudich, J. L., and Doolittle, W. F. (2006). Microbial rhodopsins: functional versatility and genetic mobility. *Trends Microbiol.* 14, 463–469. doi: 10.1016/j.tim.2006.09.006

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7. doi: 10.1038/msb.2011.75

Skennerton, C. T., Imelfort, M., and Tyson, G. W. (2013). Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Res.* 41, e105. doi: 10.1093/nar/gkt183

Sorek, R., Kunin, V., and Hugenholtz, P. (2008). CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat. Rev. Microbiol.* 6, 181–186. doi: 10.1038/nrmicro1793

Swithers, K. S., Senejani, A. G., Fournier, G. P., and Gogarten, J. P. (2009). Conservation of intron and intein insertion sites: implications for life histories of parasitic genetic elements. *BMC Evolutionary Biology* 9:303. doi: 10.1186/1471-2148-9-303

Swithers, K. S., Soucy, S. M., Lasek-Nesselquist, E., Lapierre, P., and Gogarten, J. P. (2013). Distribution and evolution of the mobile *vma-1*b intein. *Mol. Biol. Evol.* 30, 2676–2687. doi: 10.1093/molbev/mst164

Touchon, M., and Rocha, E. P. C. (2010). The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and *Salmonella*. *PLoS ONE* 5:e11126. doi: 10.1371/journal.pone.0011126

Tritt, A., Eisen, J. A., Facciotti, M. T., and Darling, A. E. (2012). An integrated pipeline for *de novo* assembly of microbial genomes. *PLoS ONE* 7:e42304. doi: 10.1371/journal.pone.0042304

Ventosa, A., Gutiérrez, M. C., Kamekura, M., Zvyagintseva, I. S., and Oren, A. (2004). Taxonomic study of *Halorubrum distributum* and proposal of *Halorubrum terrestre* sp. nov. *Int. J. Syst. Evol. Microbiol.* 54, 389–392. doi: 10.1099/ijs.0.02621-0

Vulič, M., Dionisio, F., Taddei, F., and Radman, M. (1997). Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc. Natl. Acad. Sci. U.S.A.* 94, 9763–9767. doi: 10.1073/pnas.94.18.9763

Whitaker, R. J., Grogan, D. W., and Taylor, J. W. (2003). Geographic barriers isolate endemic populations of hyperthermophilic Archaea. *Science* 301, 976–978. doi: 10.1126/science.1086909

Williams, D., Gogarten, J. P., and Papke, R. T. (2012). Quantifying homologous replacement of loci between haloarchaeal species. *Genome Biol. Evol.* 4, 1223–1244. doi: 10.1093/gbe/evs098

Wilson, G. G., and Murray, N. E. (1991). Restriction and modification systems. *Annu. Rev. Genet.* 25, 585–627. doi: 10.1146/annurev.ge.25.120191.003101

Xu, X.-W., Wu, Y.-H., Zhang, H., and Wu, M. (2007). *Halorubrum arcis* sp. nov., an extremely halophilic archaeon isolated from a saline lake on the Qinghai–Tibet Plateau, China. *Int. J. Syst. Evol. Microbiol.* 57, 1069–1072. doi: 10.1099/ijs.0.64921-0

Zhaxybayeva, O., Stepanauskas, R., Mohan, N. R., and Papke, R. T. (2013). Cell sorting analysis of geographically separated hypersaline environments. *Extremophiles* 17, 265–275. doi: 10.1007/s00792-013-0514-z

Zvyagintseva, I. S., and Tarasov, A. L. (1987). Extreme halophilic bacteria from saline soils. *Mikrobiologiya* 56, 839–844.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 03 January 2014; accepted: 18 March 2014; published online: 11 April 2014.
Citation: Fullmer MS, Soucy SM, Swithers KS, Makkay AM, Wheeler R, Ventosa A, Gogarten JP and Papke RT (2014) Population and genomic analysis of the genus Halorubrum. Front. Microbiol. 5:140. doi: 10.3389/fmicb.2014.00140
This article was submitted to Extreme Microbiology, a section of the journal Frontiers in Microbiology.

Chapter 2.2 Bioinformatic Genome Comparisons for Taxonomic and Phylogenetic Assignments Using *Aeromonas* as a Test Case

This section consists of a paper Sophie Colston and I co-first authored in 2014 (Colston et al., 2014). The major results of this paper include constructing well-supported species and species groups, the confirmation of multiple proposed misidentifications, and the new identification as misclassified of yet more taxa. Finally, the paper finds evidence for MLSA analyses being inappropriate for the *Aeromonas* group. Sophie performed the vast majority of the isolation and sequencing of the novel genomes. I performed the vast majority of the bioinformatics in the paper. Specifically, the MLSA, homologous group clustering, expanded-core phylogeny, ANI, *is*DDH, and AU-tests of the phylogenies. We both participated in drafting and editing the manuscript.

RESEARCH ARTICLE

Bioinformatic Genome Comparisons for Taxonomic and Phylogenetic Assignments Using *Aeromonas* as a Test Case

Sophie M. Colston,[a] Matthew S. Fullmer,[a] Lidia Beka,[a] Brigitte Lamy,[b,c] J. Peter Gogarten,[a] Joerg Graf[a]

Department of Molecular and Cell Biology, University of Connecticut, Storrs, Connecticut, USA[a]; Laboratoire de Bactériologie-Virologie, UMR 5119, Equipe Pathogènes et Environnements, Université Montpellier, Montpellier, France[b]; Laboratoire de Bactériologie, Centre Hospitalier Universitaire de Montpellier, Montpellier, France[c]

S.M.C. and M.S.F. contributed equally to this article.

ABSTRACT Prokaryotic taxonomy is the underpinning of microbiology, as it provides a framework for the proper identification and naming of organisms. The "gold standard" of bacterial species delineation is the overall genome similarity determined by DNA-DNA hybridization (DDH), a technically rigorous yet sometimes variable method that may produce inconsistent results. Improvements in next-generation sequencing have resulted in an upsurge of bacterial genome sequences and bioinformatic tools that compare genomic data, such as average nucleotide identity (ANI), correlation of tetranucleotide frequencies, and the genome-to-genome distance calculator, or *in silico* DDH (*is*DDH). Here, we evaluate ANI and *is*DDH in combination with phylogenetic studies using *Aeromonas*, a taxonomically challenging genus with many described species and several strains that were reassigned to different species as a test case. We generated improved, high-quality draft genome sequences for 33 *Aeromonas* strains and combined them with 23 publicly available genomes. ANI and *is*DDH distances were determined and compared to phylogenies from multilocus sequence analysis of housekeeping genes, ribosomal proteins, and expanded core genes. The expanded core phylogenetic analysis suggested relationships between distant *Aeromonas* clades that were inconsistent with studies using fewer genes. ANI values of ≥96% and *is*DDH values of ≥70% consistently grouped genomes originating from strains of the same species together. Our study confirmed known misidentifications, validated the recent revisions in the nomenclature, and revealed that a number of genomes deposited in GenBank are misnamed. In addition, two strains were identified that may represent novel *Aeromonas* species.

IMPORTANCE Improvements in DNA sequencing technologies have resulted in the ability to generate large numbers of high-quality draft genomes and led to a dramatic increase in the number of publically available genomes. This has allowed researchers to characterize microorganisms using genome data. Advantages of genome sequence-based classification include data and computing programs that can be readily shared, facilitating the standardization of taxonomic methodology and resolving conflicting identifications by providing greater uniformity in an overall analysis. Using *Aeromonas* as a test case, we compared and validated different approaches. Based on our analyses, we recommend cutoff values for distance measures for identifying species. Accurate species classification is critical not only to obviate the perpetuation of errors in public databases but also to ensure the validity of inferences made on the relationships among species within a genus and proper identification in clinical and veterinary diagnostic laboratories.

Received 13 October 2014 **Accepted** 17 October 2014 **Published** 18 November 2014

Citation Colston SM, Fullmer M, Beka L, Lamy B, Gogarten JP, Graf J. 2014. Bioinformatic genome comparisons for taxonomic and phylogenetic assignments using *Aeromonas* as a test case. mBio 5(6):e02136-14. doi:10.1128/mBio.02136-14.

Editor Edward G. Ruby, University of Wisconsin Madison

Address correspondence to J. Peter Gogarten, gogarten@uconn.edu, or Joerg Graf, joerg.graf@uconn.edu.

This article is a direct contribution from a Fellow of the American Academy of Microbiology.

Rapid improvements in DNA sequencing technologies are providing new approaches to address prevailing questions in the field of microbiology (1–3). For example, next-generation sequencing greatly enhanced the discovery of virulence factors through comparative genomics (4), enabled epidemiological studies of recent disease outbreaks (5), led to the discovery of the rare biosphere (6), and provided insights into the physiology of uncultured microbes through metatranscriptomics (7). The increasing amounts of data also brought challenges in ensuring the accuracy of annotations in databases (8). Since many analyses are based on comparisons to known sequences, errors in a database can be easily propagated in other databases and affect subsequent studies. Microbial taxonomy is one area in which the advances in next-generation sequencing have yet to be implemented to their full potential, even though several applications have shown great promise (9, 10). Prokaryotic taxonomy has been traditionally regarded as consisting of three interrelated components: classification, nomenclature, and characterization (11). Only nomenclature is strictly regulated in the International Code of Nomenclature of Bacteria (12). It is important to reconcile nomenclature when rigorous classification and characterization methods reveal an inconsistency in the composition of a particular named species.

The organizing principle of microbial taxonomy is to group

related organisms together that are distinct from other groups. DNA-DNA hybridization (DDH) is the traditional "gold standard" of circumscribing a bacterial species, as this method provides an assessment of the overall similarity of the heritable material, with phylogenetic data providing information about neighboring organisms. The current DDH standard for strains to be considered belonging to the same species is that ≥70% of the DNA from the two strains reassociates with a ≤5°C difference in melting temperatures (13). However, laboratory-based DDH measurements are not without challenges, given that DDH values can be difficult to reproduce and therefore may vary, depending on the reannealing temperature used or a laboratory's particular method employed (14). In addition, the data cannot be archived, nor are they portable between laboratories, and as such the data cannot be readily built upon when describing a new species (15).

In contrast to DDH, DNA sequence information can be easily archived and readily transferred between laboratories. Standardized bioinformatic analyses on the same data set can be performed by different laboratories, which facilitates collaborations and, potentially, the resolution of disagreements (16). Examples of such molecular methods include multilocus sequence analysis (MLSA), which provides important information about the evolutionary relationships of bacteria and allows grouping of related strains (14). MLSA has emerged as a powerful tool for classifying bacterial strains, as it relies on the allelic differences among multiple conserved housekeeping genes (17). In MLSA, the sequences are typically concatenated to overcome the lack of resolution seen in the topology of single-gene trees, but this method may mask the different evolutionary processes underlying the individual genes (18, 19). In addition, there is no consensus as to what degree of sequence variation correlates with species boundaries, which is partly due to different genes evolving at different rates and also that a few selected genes represent only a fraction of the vast amount of information contained within an entire genome.

The field of microbiology is undergoing dramatic changes, with more genomes becoming available due to the rapidly improving technology and declining cost of sequencing. In addition to closed or finished genomes, "improved" high-quality draft genomes for which the annotations have been validated have been deemed suitable for comparative genomic studies (20). The relative ease of producing such genomes provides new opportunities for assessing taxonomic relationships, discovering new taxa, and sharing data between researchers. As a result, new tools are being developed to make use of these data, including a bioinformatic approach for calculating the DDH. One of these, the genome-to-genome distance calculator, referred to here as *in silico* DDH (*is*-DDH), produces values that compare closely with experimentally derived DDH values (9, 21). Another method calculates the average nucleotide identity (ANI) among conserved and shared genes. The use of ANI has been proposed as a new standard for defining microbial species, and it is gaining wide acceptance (16, 22). The most current proposal recommends use of an ANI threshold of 95 to 96% along with support from tetranucleotide frequency correlation coefficient values (23, 24). Recently, a few studies combined either MLSA or the analysis of genes common to all members of a genus (core genome) with the overall similarity of the genome by using ANI for species identification (15, 25). We wanted to compare *is*DDH and ANI for species identification combined with phylogenetic approaches, using a genus with a complicated but relatively well-described phylogeny.

The genus *Aeromonas* makes for an ideal test case, because it contains a large number of species, biovars, and subspecies and its taxonomy has been the subject of much debate (26). Collectively, *Aeromonas* members are found in a number of habitats and in association with various animals, ranging from beneficial symbionts of leeches and zebrafish to pathogens of amphibians, fish, and humans (26, 27). Fourteen species of *Aeromonas* were recognized in the latest addition of *Bergey's Manual of Systematic Bacteriology* in 2005 (28). Since then, over a dozen have been propose, while the statuses of five species and two subspecies have been called into question. An accurate taxonomy for this genus is not only critical as a tool to differentiate benign from potentially virulent species, but it is also essential as the foundation for ecological studies.

A number of taxonomic controversies exist within the *Aeromonas* genus, namely, the synonymity of the following groups: the proposed novel species *A. culicicola* and *A. ichthiosmia* with *A. veronii* (29–31), *A. enteropelogenes* with *A. trota* (31–34), *A. allosaccharophila* with *A. veronii* (30), *A. hydrophila* subsp. *anaerogenes* with *A. caviae* (28, 35), and *A. hydrophila* subsp. *dhakensis* with *A. aquariorum*, which ultimately led to a proposal of a new species, *A. dhakensis* (36–38), All of these controversies are likely due, at least in part, to the limitations of past and current methods to consistently distinguish to the species level. Some of these controversies (e.g., whether the taxon *A. allosaccharophila* reaches the species level) could not even be unambiguously clarified with the most recent methods, with several MLSA schemes with partial sequences of up to seven housekeeping genes (33, 34, 39–41). A finding of some of these studies and of a study investigating discrepancies in the analysis of 16S rRNA genes (42, 43) was that recombination occur frequently between members of this genus, which renders phylogenies with single or a few genes challenging.

The use of whole genome sequences has been regarded as a promising avenue for the future of *Aeromonas* taxonomic and phylogenetic studies (41). In the present study, we generated improved, high-quality draft genome sequences from 27 type strains and 6 additional strains. These genomes were supplemented by 23 additional genomes of *Aeromonas* strains available in public databases. Our approach was to determine the phylogeny in three ways, by using (i) 16 housekeeping genes that were used in four recent MLSA classifications (HK), (ii) ribosomal protein coding gene (RG), and (iii) the expanded core (EC), which are the genes present in at least 90% of the 56 strains. In addition, we performed ANI analysis and *is*DDH (9, 16, 21, 22) to determine the overall similarity of the genomes. We examined our data with regard to the above-mentioned taxonomic controversies, as these provided the means to validate our approach. We also investigated the relationships of deeper phylogenetic branches in the *Aeromonas* genus. This approach led to the identification of candidate novel species and is presented as a methodology that may be applied to other genera as well.

RESULTS

Genome sequences. A total of 56 *Aeromonas* genomes were used in this study, representing type strains of 29 currently recognized or proposed species, of which 27 were sequenced in-house and 2 were available in GenBank. The additional 23 genomes were nontype strains and auxiliary strains of interest. For seven of the *Aeromonas* species, multiple strains were used in this study, and strain designations were employed to distinguish among them (*A. allosaccharophila*, *A. caviae*, *A. dhakensis*, *A. hydrophila*, *A. media*,

A. salmonicida, and *A. veronii*); for the remainder of the species, only the type strain was used, which is indicated by a superscript T. For the 33 genomes obtained for this study, the average genome coverage ranged from 30- to 260-fold and the number of scaffolds ranged from 22 to 332 with an average of 88 (Table 1). The completeness of the genomes was assessed by screening the genomes for 16 housekeeping genes and 47 ribosomal protein-coding genes. All 63 genes were present in the 56 genomes. The genome sizes estimated from the draft genomes generated for this study ranged from 3.90 Mbp (*A. fluvialis*^{T}) to 5.18 Mbp (*A. piscicola*^{T}), with an average of 4.51 Mbp. The average G+C content of the aeromonads ranged from 58.1% (*A. australiensis*^{T}) to 62.8% (*A. taiwanensis*^{T}), with a mean of 60.2%. Based on the quality of the genomes and verification of the automated annotation, we consider these genomes to be improved, high-quality draft genomes (20).

Phylogenetic analysis. One goal of our study was to reevaluate the phylogenetic relationships of the *Aeromonas* species by using three phylogenies, HK, RP, and EC, derived from different sets of genes: 16 housekeeping genes, 47 ribosomal protein-coding genes, and the expanded core, which included 2,710 ortholog groups (OG), respectively. Due to the differences in the number of informative sites, the EC phylogeny had the strongest support values for all of the nodes, although both the HK and EC phylogenies provided new insights into the relationships of distant clades (Fig. 1). The RP phylogeny had the lowest support values, as these genes are more conserved (see Fig. S1 in the supplemental material). In both the HK and EC phylogenies, we found the same eight major monophyletic groups, or clades, which are defined as groups of taxa in a phylogeny that each share an ancestor, to the exclusion of all other taxa included in the analysis (Fig. 1). Interestingly, we found several differences between the HK and EC phylogenies. In the HK phylogeny, clades 6 and 7 represent shallow branches that are nested within larger groups formed by clades 2 to 7, 3 to 7, and 4 to 7; however, in the EC phylogeny, clade 6 is basal to the large clade containing clades 2 to 5, 7, and 8. Moreover, in the EC phylogeny, clades 2 and 7 form one clade, while clades 3 to 5 form another clade, which is also inconsistent with the HK phylogeny where clade 7 forms a clade with 6 that is nested within a large grouping containing clades 3 to 7. As the expanded core did not require each ortholog group (i.e., homologs that appear to have evolved from the same ancestral gene in the organismal most recent ancestor of the group) to be present in every genome, we repeated the analysis using the strict core with only those ortholog groups that were present in all genomes. The strict core phylogeny was consistent with the EC phylogeny (see Fig. S2 in the supplemental material), indicating that the ortholog groups present in all genomes did not represent variations in the topology observed between the strict versus expanded core.

Most of the general relationships observed in our study were consistent with those reported in the published literature. The recently proposed species, *A. dhakensis*, which was determined to be synonymous with *A. aquariorum* (44), was originally a subspecies of *A. hydrophila*. All three phylogenies support that these strains form one well-supported clade that is distinct from *A. hydrophila*. Interestingly, six *A. hydrophila* genomes that we obtained from GenBank clearly clustered within *A. dhakensis*. Our study also grouped the strain SSU with *A. dhakensis*, which supports its recent reclassification from *A. hydrophila* to *A. dhakensis* (45). Misnamed genomes in GenBank should be corrected and resolved with thorough classification data to prevent further misidentifications.

Our comprehensive analysis revealed an important difference compared to the previous MLSA by Murcia-Martinez et al., which was based on partial sequences of seven genes (34). In that study, the *A. trota* isolates (which included *A. enteropelogenes*^{T}) grouped with *A. hydrophila* and *A. aquariorum*, whereas in the HK and EC phylogenies of our study, *A. enteropelogenes*^{T} and *A. trota*^{T} formed a clade with a group that included the *A. veronii* group, or AVG (*A. veronii* bv. sobria, *A. veronii* bv. veronii, and *A. allosaccharophila*), and *A. jandaei*^{T}. This finding is in agreement with those of the study by Roger et al. (33). Examination of individual gene trees suggests that the varied placement was due to the use of different housekeeping genes in these two studies (see Fig. S3 to S6 in the supplemental material) and underscores the limitations of MLSA approaches that use shorter fragments of fewer genes, compared to studies using the expanded core or a large set of full-length housekeeping genes. Our study also confirmed the synonymity of *A. trota* and *A. enteropelogenes* (31, 32).

The AVG itself is a controversial collection of species, which includes *A. culicicola*^{T} and *A. ichthiosmia*^{T}, both initially described as new species but subsequently reclassified as *A. veronii* based on DNA relatedness and biochemical characterization (29–31). Our data support the synonymity of *A. culicicola*^{T} and *A. ichthiosmia*^{T} with *A. veronii*, as the two strains grouped together with the *A. veronii* strains in one well-supported clade (Fig. 1A and B). An interesting aspect of this species is that there are two reported *A. veronii* biovars, which differ phenotypically in that *A. veronii* bv. veronii is positive (100%) for esculin hydrolysis and ornithine decarboxylation while *A. veronii* bv. sobria is negative for both reactions (46). In our analysis, the three strains of *A. veronii* bv. veronii (CECT 4257^T, AMC35, AER397) grouped together with *A. veronii* B565 in a strongly supported clade within the larger *A. veronii* clade, which supports *A. veronii* bv. veronii as a bona fide biovar. Comparisons of the *A. veronii* genomes revealed that members of *A. veronii* bv. veronii encode a β-glucosidase (EC 3.2.1.21; 793 aa) and an ornithine decarboxylase (EC 4.1.1.17; 745 aa) not found among members of *A. veronii* bv. sobria, suggesting that these two enzymes may facilitate the reactions involving esculin and ornithine, respectively. Based on this data, *A. veronii* B565, whose genome contains both genes, is a presumptive member of the *A. veronii* bv. veronii. The two *A. allosaccharophila* strains (CECT 4199^T and BVH88) also formed a strongly supported clade that was near but distinct from *A. veronii*, which suggests that *A. allosaccharophila* is a separate species. In our analysis, we also included the newest proposed *Aeromonas* species, *A. australiensis*^{T}, which is monophyletic with *A. fluvialis*^{T} and *A. sobria*^{T} and the AVG.

The other phylogenetic relationships supported the relationships described in previously published reports, such as the well-supported clade formed by *A. simiae*^{T}, *A. diversa*^{T}, and *A. schubertii*^{T} that is distinct from all the other *Aeromonas* species (Fig. 1) and observed in all three phylogenies. The close relatedness between *A. piscicola* and *A. bestiarum* (47) was also recovered in our analyses. Our results also support that strain CECT 4221, described as *A. hydrophila* subsp. *anaerogenes*, clusters within the *A. caviae* taxon.

TABLE 1 General features of the *Aeromonas* genomes

Species	Strain	Genome size (Mbp)	No. of scaffolds	Avg genome coverage[e]	N_{50}[f] (nt)	G+C content (%)	No. of predicted CDSs[g]	Accession no.	Reference
A. allosaccharophila	CECT 4199^T	4.66	120	87	114,541	58.4	4,173	PRJEB7019[a]	This study
A. dhakensis {*A. aquariorum*}[b]	CECT 7289 ^{T}	4.69	78	117	163,504	61.7	4,266	PRJEB7020[a]	This study
A. australiensis	CECT 8023^T	4.11	113	128	95,095	58.1	3,733	PRJEB7021[a]	This study
A. bestiarum	CECT 4227^T	4.68	41	53	237,067	60.5	4,223	PRJEB7022[a]	This study
A. bivalvium	CECT 7113^T	4.28	69	30	149,050	62.3	3,909	PRJEB7023[a]	This study
A. caviae	CECT 838 ^{T}	4.47	111	95	101,663	61.6	4,081	PRJEB7024[a]	This study
A. culicicola	CIP 107763^T	4.43	64	87	188,049	58.9	4,012	PRJEB7047[a]	This study
A. diversa	CECT 4254^T	4.06	37	116	203,531	61.5	3,711	PRJEB7026[a]	This study
A. encheleia	CECT 4342^T	4.47	35	112	380,984	61.9	4,076	PRJEB7027[a]	This study
A. enteropelogenes	CECT 4487^T	4.47	46	56	208,775	59.5	4,054	PRJEB7028[a]	This study
A. eucrenophila	CECT 4224^T	4.54	22	50	441,212	61.1	4,113	PRJEB7029[a]	This study
A. fluvialis	LMG 24681^T	3.90	76	48	108,949	58.2	3,609	PRJEB7030[a]	This study
A. ichthiosmia	CECT 4486^T	4.41	66	70	147,024	58.4	3,997	PRJEB7050[a]	This study
A. jandaei	CECT 4228^T	4.50	58	55	161,393	58.7	4,065	PRJEB7031[a]	This study
A. hydrophila subsp. *hydrophila*	CECT 839^T	4.74	1	UNK[c]	4,744,448	61.5	4,119	CP000462[d]	74
A. media	CECT 4232^T	4.48	233	60	37,608	60.9	4,075	PRJEB7032[a]	This study
A. molluscorum	CIP 108876^T	4.23	309	9	21,565	59.2	3,946	AQGQ01[d]	75
A. piscicola	LMG 24783^T	5.18	91	99	150,424	59.0	4,713	PRJEB7033[a]	This study
A. popoffii	CIP 105493^T	4.76	105	67	113,495	58.4	4,331	PRJEB7034[a]	This study
A. rivuli	DSM 22539^T	4.53	102	99	155,151	60.0	4,149	PRJEB7035[a]	This study
A. salmonicida subsp. *salmonicida*	CIP 103209^T	4.74	128	117	89, 543	58.5	4,442	PRJEB7036[a]	This study
A. sanarellii	LMG 24682^T	4.19	98	121	82,664	63.1	3,828	PRJEB7037[a]	This study
A. schubertii	CECT 4240^T	4.13	111	260	108,810	61.7	3,808	PRJEB7038[a]	This study
A. simiae	CIP 107798^T	3.99	100	86	73,112	61.1	3,654	PRJEB7039[a]	This study
A. sobria	CECT 4245^T	4.68	52	34	188,072	58.6	4,160	PRJEB7040[a]	This study
A. taiwanensis	LMG 24683^T	4.24	106	66	85,294	62.8	3,884	PRJEB7041[a]	This study
A. tecta	CECT 7082^T	4.76	51	89	238,229	60.1	4,278	PRJEB7042[a]	This study
A. trota	CECT 4255^T	4.34	27	66	640,249	60.0	3,917	PRJEB7043[a]	This study
A. veronii bv. veronii	CECT 4257^T	4.52	52	59	181,171	58.8	4,070	PRJEB7044[a]	This study
A. allosaccharophila	BVH88	4.71	131	204	74,486	58.6	4,295	PRJEB7045[a]	This study
A. caviae	Ae398	4.44	149	UNK	76,364	61.4	3,866	CACP01[d]	76
A. caviae {*A. hydrophila* subsp. *anaerogenes*}	CECT 4221	4.58	332	66	31,465	61.0	4,207	PRJEB7046[a]	This study
A. dhakensis {*A. aquariorum*}	AAK1	4.77	37	20	404,457	61.7	4,237	PRJDB70[d]	77
A. dhakensis {*A. hydrophila* subsp. *dhakensis*}	CIP 107500	4.71	73	84	165,885	61.8	4,284	PRJEB7048[a]	This study
A. dhakensis {*A. hydrophila*}	173	4.79	74	46	119,625	61.6	4,134	AOBN01[d]	78
A. dhakensis {*A. hydrophila*}	277	4.79	41	76	282,384	61.6	4,213	AOBQ01[d]	78
A. dhakensis {*A. hydrophila*}	14	4.67	75	45	130,840	62	UNK	AOBM01[d]	78
A. dhakensis {*A. hydrophila*}	116	4.61	45	66	208,249	62	4,090	ANPN01[d]	78
A. dhakensis {*A. hydrophila*}	259	4.70	80	39	117,245	61.7	4,098	AOBP01[d]	78
A. dhakensis {*A. hydrophila*}	187	4.78	59	111	197,352	61.6	4,205	AOBO01[d]	78
A. dhakensis {*A. hydrophila*}	SSU	4.94	2	285	4,791,870	61.5	4,449	AGWR01[d]	The Broad Institute
A. hydrophila	ML09_119	5.02	UNK	UNK	UNK	60.8	4,434	CP005966.1[d]	79
A. hydrophila	SNUFPC_A8	4.97	41	37	234,812	60.8	4,352	AMQA01[d]	80
A. hydophila subsp. *ranae*	CIP 107985	4.68	107	140	90,304	61.6	4,268	PRJEB7049[a]	This study
A. media	WS	4.78	1	UNK	4,788,430	60.7	4,385	CP007567.1[d]	81

(Continued on following page)

TABLE 1 (Continued)

Species	Strain	Genome size (Mbp)	No. of scaffolds	Avg genome coverage[e]	N_{50}[f] (nt)	G+C content (%)	No. of predicted CDSs[g]	Accession no.	Reference
A. salmonicida subsp. *achromogenes*	AS03	4.96	69	21	124,543	58.3	UNK	AMQG02[d]	82
A. salmonicida subsp. *salmonicida*	A449	5.04	1	UNK	5,040,536	58.2	4,436	CP000644.1[d]	83
A. salmonicida subsp. *salmonicida*	01-B526	4.92	604	40	83,743	58.4	4,529	AGVO01[d]	84
Aeromonas sp. {*A. hydrophila*}	AH4	4.87	41	90	258,555	59.6	4,453	PRJEB6940[a]	This study
Aeromonas sp. {*A. veronii*}	AMC 34	4.58	1	288	4,578,728	58.5	4,117	AGWU01[d]	The Broad Institute
A. veronii	B565	4.55	1	UNK	4,551,783	58.7	4,073	CP002607[d]	85
A. veronii bv. sobria	AER 39	4.42	4	283	1,516,045	58.9	3,948	AGWT01[d]	The Broad Institute
A. veronii bv. sobria	Hm21	4.68	50	200	179,631	58.7	4,245	ATFB01[d]	62
A. veronii bv. sobria	LMG 13067	4.74	72	46	147,470	58.3	4,171	PRJEB7051[a]	This study
A. veronii bv. veronii	AER 397	4.50	5	378	3,260,625	58.9	3,986	AGWV01[d]	The Broad Institute
A. veronii bv. veronii	AMC 35	4.57	2	285	4,172,420	58.6	4,036	AGWW01[d]	The Broad Institute

[a] Obtained from the EMBL Nucleotide Sequence Database.
[b] Previously published names are indicated inside braces.
[c] UNK, unknown.
[d] Obtained from GenBank, National Center for Biotechnology Information.
[e] The average genome coverage is expressed in bp sequenced divided by genome size.
[f] The N_{50} (reported in nucleotides) represents the smallest of the largest contigs covering 50% of the total size of all contigs.
[g] CDS, coding sequence.

Assessment of genome similarity using *is*DDH and ANI. The information gained from the phylogenetic analyses provides an important depiction of the evolutionary relationships of different strains but does not translate directly into the overall similarity of the genomes, which was determined through DDH. We used two different *in silico* or bioinformatics approaches, *is*DDH and ANI, that have been proposed to overcome the challenges of conventional laboratory-based DDH to evaluate the genomic similarity of bacteria, and we evaluated the congruence of these methods (Fig. 2) (9, 16, 21, 22).

Two excellent examples for validating this approach are *A. culicicola*^{T} and *A. ichthiosmia*^{T}, which were initially proposed as novel species and later reclassified as *A. veronii* based in part on DDH values that exceeded 70%. The predicted point estimates of the *is*DDH values we obtained for these two strains were all slightly below 70% (69.1 to 69.6% and 67.4 to 68.2, respectively) compared to all other named *A. veronii* strains (see Fig. S7 in the supplemental material). However, when taking into consideration the 95% confidence interval (CI) for every comparison of these two strains, all CIs encompassed the 70% threshold (upper CI borders, 70.6 to 71.8%), affirming that they are indeed *A. veronii*. While these *is*DDH values were lower than what we observed for other pairwise *A. veronii* strain comparisons, the median hybridization value for *A. culicicola*^{T} and *A. ichthiosmia*^{T} to *A. veronii* was only 2.2% below that of the *A. veronii* comparisons (71.6% versus 73.8%). Additionally, both strains also had ANI values at or above the 96% level, compared to the other named *A. veronii* strains, which supports that *A. culicicola*^{T} and *A. ichthiosmia*^{T} are part of the *A. veronii* species, albeit near the periphery. The *is*DDH and ANI values were consistent with previously published results (29, 30).

The taxonomic status of *A. allosaccharophila* has been controversial, and it has been suggested that it is a member of *A. veronii* (30). The upper borders of the 95% CI for the *is*DDH values for *A. allosaccharophila* are below 70% compared to the *A. veronii* strains. Additionally, the ANI values are all ~94%. These data support the status of *A. allosaccharophila* as a bona fide species that is closely related to *A. veronii*. Interestingly, while the HK, RP, and EC phylogenies all grouped the two *A. allosaccharophila* genomes (CECT 4199^T and BVH88) together and separate from *A. veronii*, the ANI and the upper 95% CI *is*DDH values between the two *A. allosaccharophila* genomes were both just under the species cutoff boundary, at 95.8% and 68.7%, respectively. These data suggest that BVH88 may not be a member of the *A. allosaccharophila* species, but a greater number of strains in this clade will need to be evaluated to clarify their relationships. Two other species, *A. fluvialis* (ANI, ~92%) and *A. australiensis* (ANI, ~93%), also group near *A. veronii*. Their *is*DDH estimates register ~52% compared to *A. veronii*.

Another group of species that has recently attracted attention is *A. aquariorum*, *A. hydrophila* subsp. *dhakensi*, and *A. hydrophila*. The partition of the group comprised of *A. aquariorum*/*A. hydrophila* subsp. *dhakensis* strains from the *A. hydrophila* group, which includes the type strain (CECT 839), was recovered conclusively by every method we used in our study. The branch lengths of the HK phylogeny between *A. dhakensis* and *A. hydrophila* (~0.075 substitutions/site) were similar to those separating many named species in the HK reconstruction, such as those between *A. eucrenophila*^{T} and *A. tecta*^{T} (~1.0 substitutions/site), *A. schubertii*^{T} and *A. diversa*^{T} (~0.09 substitutions/site), *A. rivuli*^{T} and *A. molluscorum*^{T} (~0.06 substitutions/site), and *A. piscicola*^{T} and *A. bestiarum*^{T} (~0.04 substitutions/site). Similar relationships were observed in the RP and EC phylogenies. Further evidence comes from the ANI data, which showed only 93% similarity between the two different clades. This is well below the 96% species cutoff recommended by Richter (23). This conclusion was further sup-

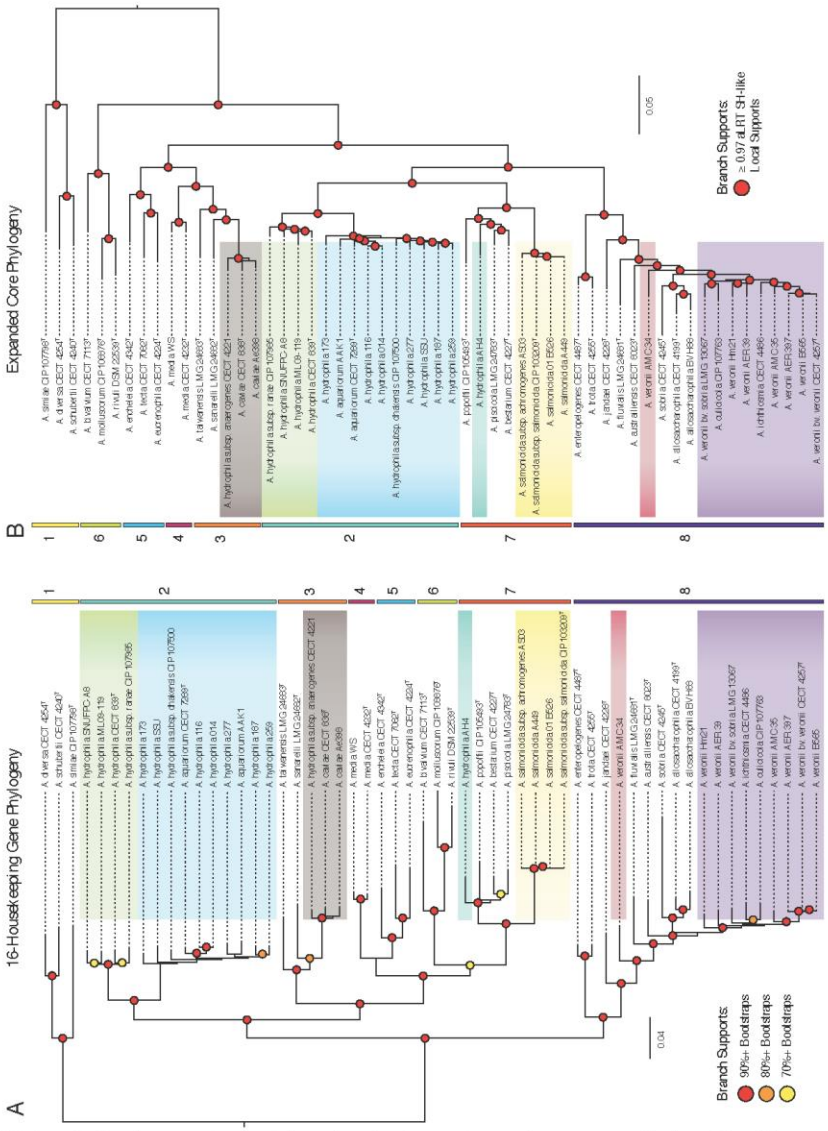


FIG 1 (A) Maximum likelihood reconstruction of 16 single-copy housekeeping genes. Support values are represented by dots: red (90% bootstraps), orange (80%), yellow (70%). (B) Approximate maximum likelihood reconstruction of 2,710 orthologous groups found in 90% or more of the taxa. aLRT SH-like support values equal to or greater than 0.97 are represented by red dots. The species *A. veronii*, *A. hydrophila*, *A. dhakensis*, *A. salmonicida*, and *A. caviae* are color-coded in both trees. Additionally, two previously misidentified taxa, *A. veronii* AMC 34 and *A. hydrophila* AH4, are shown in red and teal, respectively. Eight well-supported clades were shared between the two reconstructions. They are shown by the colored bars and are numbered 1 through 8.

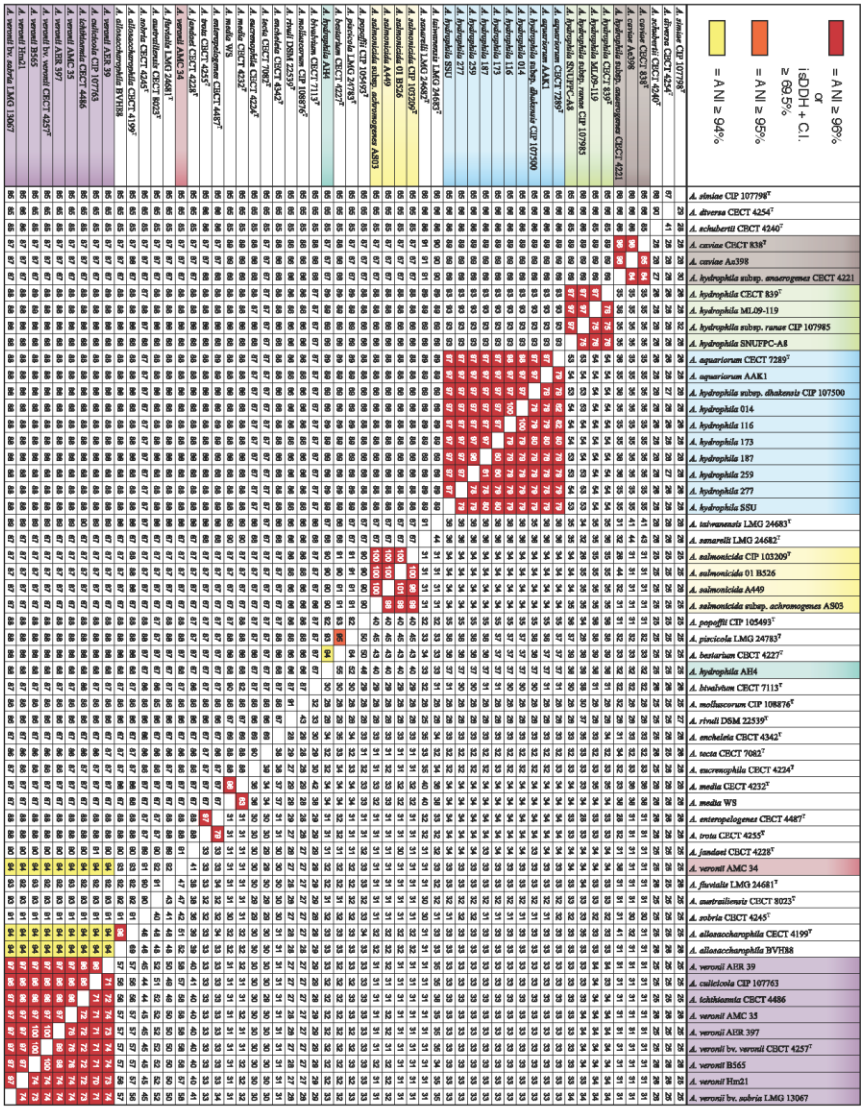


FIG 2 ANI and *is*DDH values. The lower triangle displays ANI values, and the upper triangle shows the *is*DDH values. ANI values are colored according to three historical species cutoff values: 94% (yellow), 95% (orange), and 96% (red). The *is*DDH values displayed are the upper limits of the 95% confidence intervals and are colored red if the met the laboratory DDH species cutoff of 70% hybridization. ANI of 96% correlates well with 70% *is*DDH values, with only the *A. allosaccharophila* isolates failing to match (68.7%).

ported by *is*DDH data, in which *A. dhakensis* and *A. hydrophila* strains all scored below 60% between species when using the upper border of the 95% CI, while within each partition all values were well above 70%. These data confirm that these two clades represent two discrete species rather than constituents of one, as was originally proposed (48).

A. piscicola^T and *A. bestiarum*^{T} grouped together and formed one clade with *A. popoffii*^{T}. The ANI between *A. piscicola*^{T} and *A. bestiarum*^{T} was 95.2%, which is near the 96% suggested species cutoff (23). However, while their *is*DDH values were higher than most between-species comparisons (61.1% point estimate, 64.4% at the upper 95% CI), they still fell short of what one would expect for members of the same species. It will be important to add more strains of these two groups in future analyses to gain better insight into the relationships between these taxa. Based on the current data, a 96% cutoff for the ANI value seems appropriate for *Aeromonas* species delineations.

Discovery of novel species. We also included two strains in our analysis that seemed unusual based either on previous studies or preliminary data. AMC 34, a clinical isolate described as *A. veronii* bv. veronii, had a long branch length and clustered away from other *A. veronii* bv. veronii strains in a previous study (41). Strain AH4 was published as *A. hydrophila* by investigators that had obtained this isolate from the water of a storage container for medicinal leeches (49). In the HK phylogeny, AMC 34 clustered well outside the *A. veronii* clade, near *A. jandaei* ^{T}and *A. fluvialis*^{T}, with bootstrap support values in excess of 90% (Fig. 1A). Similarly, the EC phylogeny placed AMC 34 outside of *A. veronii* with high support (Fig. 1B). The ANI between AMC 34 and the AVG was ~94%, while the *is*DDH was only ~58% compared to the same taxa (Fig. 2). Taken together, the data strongly support AMC 34 as a new species.

The other strain, AH4, was identified by a clinical diagnostic laboratory as *A. hydrophila* (49). In all of our phylogenetic analyses, AH4 grouped with *A. piscicola*^{T} and *A. bestiarum*^{T} with high support. This placement and its distance from *A. hydrophila* were strongly supported by the ANI and *is*DDH data (Fig. 2). AH4 registered only ~89% to both the *A. hydrophila* and *A. dhakensis* groups but much higher values to *A. bestiarum*^{T} (~94%) and *A. piscicola*^{T} (~93%). *is*DDH also supported the conclusion that AH4 is not likely a member of *A. bestiarum* (~55%) or *A. piscicola* (~52%) and is distinct from the *A. hydrophila* (~38%) and *A. dhakensis* (37%) groups.

All of our bioinformatics analyses indicated that the strains AMC 34 and AH4 represent two new species; however, we were restricted to a single isolate of each, which precluded the assessment of the variabilities of biochemical tests (see Table S1 in the supplemental material). In addition, we were unable to include one recently published type strain, *A. cavernicola* CCM7641^T (50) or one proposed new species, *A. lusitana* (34), which has not yet been officially described. Using the available MLSA data, we were able to show that AMC 34 and AH4 did not cluster near these two species and are thus not likely members of either *A. cavernicola* or *A. lusitana* (see Fig. S8 in the supplemental material). The accessibility of the genomes published for this study will provide other researchers with the opportunity to determine the probable taxonomic position of candidate novel species, an important capability in light of the number of taxonomic problems described for *Aeromonas*.

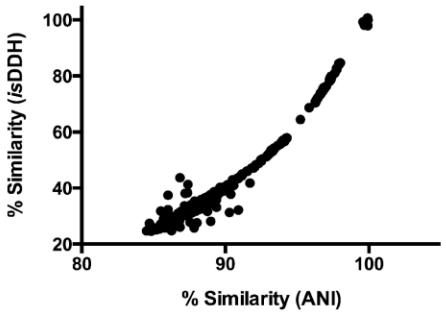


FIG 3 Comparison of *is*DDH and ANI results. The pairwise percent similarities of 56 genomes were determined using either *is*DDH or ANI. The two approaches revealed a significant correlation, with an r^2 of 0.957. When testing samples with an *is*DDH values of ≥50%, the r^2 was 0.9996.

Comparison of phylogenetic and genetic distance measures. The delineation of organisms into taxonomic groups is based on their evolutionary histories and genetic distances. In this study, we utilized five different approaches, of which two were phylogeny independent (*is*DDH and ANI) and three had a phylogenetic component (HK, RP, and EC phylogenies). To guide subsequent studies, we wanted to evaluate whether these approaches were in agreement with one another and whether some were more informative than others. Even though *is*DDH and ANI use different algorithms for the calculations, e.g., ANI evaluates the similarity of shared elements between two genomes, while *is*DDH estimates the overall similarity of two genomes, the results were very consistent (Fig. 3). The r^2 value was 0.957 for the entire data set, and when restricted to comparisons of more closely related strains (*is*DDH of ≥55%), the r^2 was 0.996. These values demonstrated that at least for this data set, either method can be used for determining overall genome similarities. When *is*DDH (upper 95% CI) and ANI results were compared to the P-distance of the entire EC data set, the r^2 values were low for both approaches, 0.599 and 0.713, respectively. When the data set was restricted to comparisons of genomes that had at least a similarity of ≥50% based on *is*DDH, the correlation coefficients were 0.943 and 0.965for *is*DDH and ANI, respectively (see Fig. S9 in the supplemental material). This indicated that either approach works well at separating closely related genomes but not for determining more distant relationships.

Most researchers characterize strains by analyzing the sequences of only one or two genes. We wanted to ascertain whether there are particular genes that are better suited than others for an initial analysis. One important concern is that horizontal gene transfer of gene fragments and not just entire genes can occur among aeromonads and result in conflicting phylogenies (41). Thus, relying on any one gene can produce erroneous results. On the other hand, including a preponderance of genes that represent a highway of gene sharing in a concatenation may result in phylogenies that reflect neither organismal evolution nor any individual gene history (51). The individual gene trees (see Fig. S3 to 6 in the supplemental material) for the 16 housekeeping genes were compared to the phylogeny derived from the consensus tree using

the approximately unbiased (AU) test (52). The set of maximum likelihood (ML) trees generated from bootstrap samples of the MLSA data were significantly different from the best gene tree for each gene. When maximum likelihood trees from bootstrap samples of the 16 housekeeping genes were compared to the MLSA tree, 15 of the gene tree sets were significantly different from the MLSA best tree. Only one of the bootstrap samples for *recA* had a *P* value of ≥0.05 ($P \geq 0.93$). These results reveal that no individual gene tree properly reflects or is even compatible with the phylogeny of the MLSA tree.

DISCUSSION

Our polyphasic genome comparison utilizing both phylogenetic and genetic distance metrics was by and large consistent with the current understanding of the phylogenetic relationships of the species contained within the genus *Aeromonas*, which had been hitherto based on laboratory-determined DDH values, biochemical tests, and multilocus sequence typing. Importantly, we were able to gain new insights into the overall relationships of the *Aeromonas* species with the phylogeny generated from the expanded core and the HK genes. There were eight major clades from the EC that were largely consistent with the HK phylogeny (Fig. 1). One major difference between the two phylogenies was the placement of *A. salmonicida* (clade 7) and *A. hydrophila* and *A. dhakensis* (clade 2). In the EC phylogeny, they form one strongly supported clade, but in the HK phylogeny they are separated by two well-supported nodes (Fig. 1). This suggests that other components of the genome are forcing *A. hydrophila* and *A. salmonicida* together in the expanded core phylogeny. Due to the limited resolution, the RP phylogeny did not provide additional support. A strict core phylogeny using only ortholog groups present in all 56 taxa shared the topology of the EC tree, suggesting that the conflict with the HK method was due to genes present in 100% of the genomes (see Fig. S2 in the supplemental material). One should consider, however, that the EC phylogeny may have inherent biases which might lead to an inaccurate depiction of organismal phylogeny. At this point, we cannot establish which topology is correct, since gene transfer between divergent groups has the potential to lead to trees from concatenated data sets that do not reflect the vertical inheritance (19). Gene transfer frequency is usually biased toward close relatives, thus reinforcing the signal due to shared ancestry (53, 54). In contrast, highways of gene sharing between more distant species can obscure the vertical phylogenetic signal due to shared ancestry (51, 55). For phylogenetic relationships within each of the clades 1 through 7, the HK and EC phylogenies appear to approximate organismal phylogeny (Fig. 1). On the other hand, relationships between these clades remain ambiguous. Differences in substitution rates and saturation with substitutions make it difficult to apply ANI and *is*DDH to higher taxonomic levels. Future work will need to include the evaluation of the 2,710 individual trees from the EC analysis in a combined analysis, such as the one described by Bansal, Alm, and Kellis (56), to determine the major conflicting phylogenetic signals retained in these genomes. Even so, both the HK and EC phylogenies provided more information regarding the relationships of different *Aeromonas* species than previous MLSA studies.

The psychrophilic aeromonads have been differentiated from the mesophilic strains based on growth physiology, biochemical properties, and virulence characteristics. Although there certainly are important differences among these characteristics, whole-genome information groups them clearly among the mesophilic species, near *A. hydrophila* and *A. dhakensis*. One interesting distinction of the *A. salmonicida* clade is that there is much less genetic diversity, indicated by the *is*DDH values for strains of the same species. The four *A. salmonicida* genomes had *is*DDH values ≥98.5%, in comparison to *A. hydrophila* (≥75.7%), *A. dhakensis* (≥78.3%), and *A. veronii* (≥70.4%). This was consistent with a study that suggested a clonal distribution of *A. salmonicida* subsp. *salmonicida* based on identical pulse electrophoresis DNA fingerprints, which showed identical banding patterns from strains isolated from different geographical regions (57). This difference in genetic diversity could reflect different evolutionary driving forces for *A. salmonicida* strains. One conjecture is that perhaps they are adapted for a virulent lifestyle in fish, where clonal outbreaks are more likely to occur. It is also possible that there is a sampling bias, which future studies employing more strains should help to resolve.

One of our goals was to assess the utility of bioinformatics approaches to replace traditional taxonomic approaches for species identification. Despite the shortcomings and challenges of laboratory DDH, whole-genome content comparisons collectively represent the most valuable criterion for demarcation of bacterial species. As more bacterial genomes are sequenced and the information is made accessible, the use of whole genome sequences in the characterization of bacterial species provides opportunities that should not be ignored. This approach has been used in clarifying the taxonomic positions in some cases, e.g., for *Acinetobacter* using ANI and core gene phylogeny (15) and for *Vibrio* using MLSA based on genome information (58). To our knowledge, however, an approach utilizing *is*DDH and ANI combined with HK, RG, and EC phylogenies has not yet been done for a genus characterized by a complicated taxonomy and using a plurality of its members.

Aeromonas is an interesting test case for a number of reasons. This genus is comprised of a large number of species capable of diverse associations depending on the species. The spectrum encompasses benign and virulent species, a range that can also exist within a single species. *A. hydrophila*, *A. caviae*, and *A. veronii* have long been associated with human disease (26). Recently, *A. dhakensis* was recognized as a new virulent species (59), a distinction obfuscated in part due to *A. dhakensis* strains initially regarded as *A. hydrophila*. Of the numerous *Aeromonas* species that have been proposed and characterized, many of those species have been redefined and renamed as new information has been presented. This shifting nomenclature is a manifestation of the inefficiencies inherent in current taxonomic methods for *Aeromonas*. While the number of publically available *Aeromonas* genomes has increased dramatically in the last few years, most of the type strains are yet to be fully sequenced. We produced improved, high-quality draft genomes for these type strains and for some non-type strains of interest. Our results recapitulated known phylogenetic relationships and provided further insights into several others. This study also identified the breakpoints between species, indicating that this approach can be used to identify new species. For demarcating species boundaries, *is*DDH and ANI produced similar results, as reflected in the correlation of the values observed when using the upper 95% CI bound to the *is*DDH estimates (Fig. 3). The current version of *is*DDH is only available in a Web-based interface that requires manually uploading the sequence information, while ANI can be easily run on local servers. Consequently, we found

ANI to be more time-effective when dealing with a large number of strains. For smaller studies, *is*DDH would be equally fast for computing and would also have the benefit of confidence intervals and probability statistics.

Apart from the fact that our approach could confidently and consistently resolve recent taxonomic controversies, our analysis also revealed that two strains, AMC 34 and AH4, represent new *Aeromonas* species. This conclusion is based on the distance in the genome content according to ANI and *is*DDH values, as well as the phylogenetic distances of the strains. These finding highlights two important advantages of bioinformatic assessment of genome similarity: (i) the expensive generation of the raw data does not have to be repeated by other research groups, and (ii) interlaboratory variations in DDH determinations can be overcome by agreeing to a cutoff value with standardized parameters in bioinformatic analyses. To facilitate the progress of other research groups in the *Aeromonas* field, we have set up a website (http://aeromonasgenomes.uconn.edu) that allows users to query and download all of the available *Aeromonas* genomes, contains the scripts we used in our analysis, and provides a summary of our current distance measures.

Another important finding from our analysis was that, out of the 23 publically available *Aeromonas* genomes that we analyzed, 8 (34.8%) are inconsistently named. In large part this was due to the recent reclassification of *A. hydrophila* subsp. *dhakensis* as *A. dhakensis* and the reclassification of *A. aquariorum* as *A. dhakensis*. While the initial misclassifications are understandable, efforts should be taken to correct and update the nomenclature to curtail the promulgation of inaccurate information. NCBI currently allows only the original submitter to request the name change (http://www.ncbi.nlm.nih.gov/books/NBK51157/). One possibility would be to involve the community at large to provide input on such discrepancies.

The ability to generate improved, high-quality draft genome sequences rapidly and inexpensively, and of a sufficient quality for robust phylogenetic analyses (20), is changing the landscape of how one can investigate microbial taxonomy and should lead to a change in the requirements of performing laboratory-based DDH for species descriptions. An additional benefit of genome sequencing is that it offers a comprehensive resource to explore the myriad of potential metabolic capabilities, physiology, virulence factors, and antibiotic resistance profiles for the strains studied. The advantages of *in silico* DDH or ANI have been elegantly stated before (9, 16, 21, 22), and we have provided strong support for implementing these approaches in today's microbial taxonomic studies. However, we recognize that the procedure of officially naming and describing new organisms is understandably a conservative and carefully regulated process; the effects on many different constituents have to be considered, since any amendments will result in broad effects for the scientific community at large. In this study, we provided data from a genus with a complex and controversial taxonomy and demonstrated the accuracy of the bioinformatics approach to identify new species and to correct erroneous identifications from previous studies. Utilizing the same software, code, and parameters for the data analysis, one can readily compare findings of other groups, thus supplanting arguments concerning laboratory methodologies with practical discussions on appropriate cutoff levels. For this test case study with *Aeromonas*, an *is*DDH of ≥70% at the upper 95% confidence interval or an ANI value of ≥96% was consistent for genomes belonging to the same species. Distance in the EC phylogeny is another metric that can be useful in species identification; in our study, a distance of ≤0.026 indicated that the genomes belong to the same species. It is likely that these types of values will also be applicable to other genera.

MATERIALS AND METHODS

Strains, growth conditions, and biochemical tests. For the genome data set, we included all of the type strains for *Aeromonas* with the exception *A. cavernicola* (50), as well as all other *Aeromonas* genomes deposited into public databases as of 17 July 2013. For the type strains, 2 were publically available and 27 were sequenced in-house. For additional strains, 21 were available publically and 6 were sequenced in-house. The bacteria were grown at the optimal growth temperature for the strain in LB broth or on LB agar (1.5%) plates for 16 to 18 h (60). For biochemical tests, API 20NE strips (bioMérieux, Marcy l'Etoile, France) were used in accordance with the manufacturer's instructions. Separate tests for ornithine decarboxylase (ODC) activity and esculin hydrolysis were assessed using ODC broth and bile esculin agar (Sigma-Aldrich, St. Louis, MO). Tests were performed in triplicate.

Library preparation and genome sequencing. Genomic DNA was extracted using the MasterPure DNA purification kit (Epicenter, Madison, WI) and quantified using a Qubit 2.0 fluorometer (Life Technologies, Carlsbad, CA). DNA was also checked for quality by using a NanoDrop instrument (NanoDrop Products, Wilmington, DE) as well as on an agarose gel. Libraries were prepared from the genomic DNA using a Nextera or Nextera XT DNA sample preparation kit (Illumina, Inc., San Diego, CA). Library concentrations were determined by using the Qubit fluorometer and bioanalyzer (Agilent Technologies, Santa Clara, CA) prior to sequencing on a MiSeq benchtop sequencer (Illumina, Inc.) at the Microbial Analysis Resources and Services facility at the University of Connecticut (Storrs, CT).

Assembly and annotation. Paired Illumina reads were trimmed and assembled into scaffolded contigs by using the *de novo* assembler of CLC Genomics Workbench versions 6.0.04 to 7.0.04 (CLC-bio, Aarhus, Denmark). Annotation of the contigs was accomplished using the Rapid Automated Annotation using Subsystem Technology (RAST) server (61). All *Aeromonas* completed and draft annotated assemblies from the NCBI ftp repository that were used in this study were downloaded, back-engineered into contigs, and submitted to RAST for reannotation to mitigate any biases in the RAST annotation algorithms by applying them equally to each genome. The completeness of the genomes was initially assessed by screening for 17 housekeeping genes and 47 ribosomal proteins. We failed to detected *ppsA* (phosphoenolpyruvate synthase) in *A. fluvialis*. A thorough investigation employing mapping of reads to reference sequences and examining the region containing *ppsA* in the other strains suggested that this gene may not be present in this organism, and thus we excluded *ppsA* from the analysis.

MLSA reference tree and individual gene tree generation. Sixteen housekeeping genes (*atpD*, *dnaJ*, *dnaK*, *dnaX*, *gltA*, *groL*, *gyrA*, *gyrB*, *metG*, *mdh*, *radA*, *recA*, *rpoC*, *rpoD*, *tsf*, and *zipA*) were used for MLSA (33, 34, 39). The DNA-directed RNA polymerase subunit beta=(*rpoC*) was used in the MLSA dataset. Adding *rpoB* to the dataset or switching it for *rpoC* did not change the phylogeny resulting from the MLSA analysis depicted in Fig. 1. These genes were initially chosen in three separate MLSA studies for their conservation among all aeromonads, ease of PCR primer design, broad distribution, and single copy number in the chromosome. The full-length sequence of each gene was initially derived from the previously published genome of *A. veronii* Hm21 (62), and these sequences served as queries for BLAST searches against the annotated proteins of all 56 genomes. Multiple sequence alignments (MSAs) were generated by translating the genes to protein sequences in SeaView (63), aligning the proteins using MUSCLE (v.3.8.31) (64) and then back-translating to the nucleotide sequences prior to the phylogenetic analysis. Each MSA was manually evaluated, and any sequences showing poor alignment were examined further, including comparison against the nonredundant data-

base using BLAST and excluded if not found to be the correct protein. In-house scripts created a concatenated alignment of all 16 genes. A model of evolution was determined by using the Akaike information criterion with correction for small sample size (AICc), as implemented in jModelTest 2.1.4. An ML phylogeny was generated from the concatenated MSA, and individual gene phylogenies from the individual gene MSAs were determined by using PhyML (v 3.0_360-500M) (65). PhyML parameters consisted of a GTR model, estimated p-invar, 4 substitution rate categories, estimated gamma distribution, and subtree pruning and regrafting enabled with 100 bootstrap replicates. Using the same approach, phylogenies were determined for each of the 16 housekeeping genes.

Ribosomal reference tree generation. Forty-seven ribosomal proteins were obtained from the BioCyc website (66). These served as queries for BLAST searches against the annotated proteins of all 56 genomes. Multiple sequence alignments were generated as described above for the MLSA tree. The AICc reported the best-fitting model to be GTR plus gamma estimation plus invariable site estimation.

Core genome comparison. To define a core genome, the annotated protein open reading frames (ORFs) from each genome were used as BLAST queries against the protein ORFs of each other genome in the study, using in-house Perl scripting. The BLAST outputs were processed into OGs with MCL-edge v14-137 (67, 68) (http://micans.org/mcl/). The inflation value was set to 10 in order to break the OGs down into smaller clusters that more closely resembled individual genes rather than families. A relaxed core was defined by extracting OGs present in at least 90% of the taxa used in this study. Where a taxon had multiple entries in a single OG, the first entry reported by MCL was arbitrarily included and the others were excluded. Each OG was aligned using MUSCLE v 3.8.31 (64). In-house Perl scripting concatenated the OGs into a single alignment. Owing to the scale of the concatenated alignment, FastTreeMP (69) was used to perform the phylogenetic reconstruction. The substitution model used was WAG.

Pairwise sequence distance calculations and identity calculations. Sequence distances were calculated using the SaveDist function in PAUP* v4.0b10 (70). The distance type calculated was the P-distance.

Average nucleotide identity/tetramer analysis. Assembled contigs were reconstituted from the RAST-generated GenBank files for all genomes by using the seqret function of the EMBOSS package (71). All genomes were treated in the same manner to ensure that any biases were consistent across the entire data set. JSpecies1.2.1 (23) was used to analyze these contig sets for the ANI and tetramer usage patterns, using default parameters. We report here the averages of the reciprocal comparisons.

Tree comparisons using the approximately unbiased test. Per site log likelihoods were generated in RAxML v 7.3.5 (72). The AU tests (52) were carried out in the CONSEL v 1.20 package (73). Comparisons were made with HK tree against the 100 bootstrap replicates from each individual gene. Likewise, each best individual gene tree was compared against 100 bootstrap replicates of the HK tree.

***In silico* DNA-DNA hybridization.** Estimates of *is*DDH were made using the Genome-to-Genome Distance Calculator (GGDC) (9, 21). The contig files were uploaded to the GGDC 2.0 Web server (http://ggdc.dsmz.de/distcalc2.php), where *is*DDH calculations were performed. Formula 2 alone was used for analysis, since it calculates *is*DDH estimates independent of genome lengths and is recommended by the authors of GGDC for use with any incomplete genomes (9, 21). The point estimate plus the 95% model-based confidence intervals were used for analysis.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.02136-14/-/DCSupplemental.

Figure S1, EPS file, 0.6 MB.
Figure S2, EPS file, 0.6 MB.
Figure S3, EPS file, 2.1 MB.
Figure S4, EPS file, 2.1 MB.
Figure S5, EPS file, 1.5 MB.
Figure S6, PDF file, 2 MB.
Figure S7, EPS file, 11.8 MB.
Figure S8, EPS file, 0.7 MB.
Figure S9, PDF file, 0.3 MB.
Table S1, DOCX file, 0.02 MB.

ACKNOWLEDGMENTS

We thank E. Talagrand for excellent technical assistance, A. Horneman and R. M. Humphries for providing strains, the UConn Bioinformatics Facility for providing computing resources and the Microbial Analysis, Resources and Services Facility for access to an Illumina MiSeq system.

This research was supported through NIH R01 GM095390 (Joerg Graf, Peter Visscher, and Hilary G. Morrison), USDA ARS agreement 58-1930-4-002, and the National Science Foundation (DEB 0830024).

REFERENCES

1. Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW. 2012. Transforming clinical microbiology with bacterial genome sequencing. Nat. Rev. Genet. 13:601–612. http://dx.doi.org/10.1038/nrm3437.
2. Pallen MJ, Loman NJ, Penn CW. 2010. High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. Curr. Opin. Microbiol. 13:625–631. http://dx.doi.org/10.1016/j.mib.2010.08.003.
3. Ribeca P, Valiente G. 2011. Computational challenges of sequence classification in microbiomic data. Brief. Bioinform. 12:614–625. http://dx.doi.org/10.1093/bib/bbr019.
4. Chen L, Xiong Z, Sun L, Yang J, Jin Q. 2012. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic Acids Res. 40:D641–D645. http://dx.doi.org/10.1093/nar/gkr989.
5. Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, Fitzgerald M, Godfrey P, Haas BJ, Murphy CI, Russ C, Sykes S, Walker BJ, Wortman JR, Young S, Zeng Q, Abouelleil A, Bochicchio J, Chauvin S, Desmet T, Gujja S, McCowan C, Montmayeur A, Steelman S, Frimodt-Møller J, Petersen AM, Struve C, Krogfelt KA, Bingen E, Weill FX, Lander ES, Nusbaum C, Birren BW, Hung DT, Hanage WP. 2012. Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. Proc. Natl. Acad. Sci. U. S. A. 109:3065–3070. http://dx.doi.org/10.1073/pnas.1121491109.
6. Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere." Proc. Natl. Acad. Sci. U. S. A. 103:12115–12120. http://dx.doi.org/10.1073/pnas.0605127103.
7. Bomar L, Maltz M, Colston S, Graf J. 2011. Directed culturing of microorganisms using metatranscriptomics. mBio 2(2):e00012-11. http://dx.doi.org/10.1128/mBio.00012-11.
8. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLOS Comput. Biol. 5:e1000605. http://dx.doi.org/10.1371/journal.pcbi.10000605.
9. Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M. 2013. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics 14:60. http://dx.doi.org/10.1186/1471-2105-14-60.
10. Chun J, Rainey FA. 2014. Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. Int. J. Syst. Evol. Microbiol. 64:316–324. http://dx.doi.org/10.1099/ijs.0.054171-0.
11. Brenner DJ, Staley JT, Krieg NR. 2005. Classification of procaryotic organisms and the concept of bacterial speciation, p 27316–32. *In* Brenner DJ, Staley JT, Krieg NR, Garrity GM (ed), Bergey's manual of systematic bacteriology, vol 2. The proteobacteria. Springer Verlag, New York, NY.
12. Lapage SP, Sneath PHA, Lessel EF, Skerman VBD, Seeliger HPR, Clark WA. 1992. International code of nomenclature of Bacteria: bacteriological code, 1990 revision. ASM Press, Washington, DC.
13. Stackebrandt E, Frederiksen W, Garrity GM, Grimont PA, Kämpfer P, Maiden MC, Nesme X, Rosselló-Mora R, Swings J, Trüper HG, Vauterin L, Ward AC, Whitman WB. 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int. J. Syst. Evol. Microbiol. 52:1043–1047. http://dx.doi.org/10.1099/ijs.0.02360-0.
14. Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Van de Peer Y, Vandamme P, Thompson FL, Swings J.

2005. Opinion: re-evaluating prokaryotic species. Nat. Rev. Microbiol. 3:733–739. http://dx.doi.org/10.1038/nrmicro1236.
15. Chan JZ, Halachev MR, Loman NJ, Constantinidou C, Pallen MJ. 2012. Defining bacterial species in the genomic era: insights from the genus *Acinetobacter*. BMC Microbiol. 12:302. http://dx.doi.org/10.1186/1471-2180-12-302.
16. Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. U. S. A. 102:2567–2572. http://dx.doi.org/10.1073/pnas.0409727102.
17. Clarke SC, Diggle MA, Edwards GF. 2002. Multilocus sequence typing and *porA* gene sequencing differentiates strains of *Neisseria meningitidis* during case clusters. Br. J. Biomed. Sci. 59:160–162.
18. Kämpfer P, Glaeser SP. 2012. Prokaryotic taxonomy in the sequencing era—the polyphasic approach revisited. Environ. Microbiol. 14:291–317. http://dx.doi.org/10.1111/j.1462-2920.2011.02615.x.
19. Lapierre P, Lasek-Nesselquist E, Gogarten JP. 2014. The impact of HGT on phylogenomic reconstruction methods. Brief. Bioinform. 15:79–90. http://dx.doi.org/10.1093/bib.bbs050.
20. Chain PSG, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, Cole JR, Ding Y, Dugan S, Field D, Garrity GM, Gibbs R, Graves T, Han CS, Harrison SH, Highlander S, Hugenholtz P, Khouri HM, Kodira CD, Kolker E, Kyrpides NC, Lang D, Lapidus A, Malfatti SA, Markowitz V, Metha T, Nelson KE, Parkhill J, Pitluck S, Qin X, Read TD, Schmutz J, Sozhamannan S, Sterk P, Strausberg RL, Sutton G, Thomson NR, Tiedje JM, Weinstock G, Wollam A, Consortium GSCHMPJ, Detter JC. 2009. Genomics. Genome project standards in a new era of sequencing. Science (New York, NY) 326:236–237. http://dx.doi.org/10.1126/science.1180614.
21. Auch AF, von Jan M, Klenk HP, Göker M. 2010. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand. Genomic Sci. 2:117–134. http://dx.doi.org/10.4056/sigs.531120.
22. Konstantinidis KT, Ramette A, Tiedje JM. 2006. Toward a more robust assessment of intraspecies diversity, using fewer genetic markers. Appl. Environ. Microbiol. 72:7286–7293. http://dx.doi.org/10.1128/AEM.01398-06.
23. Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. U. S. A. 106:19126–19131. http://dx.doi.org/10.1073/pnas.0906412106.
24. Scortichini M, Marcelletti S, Ferrante P, Firrao G. 2013. A genomic redefinition of *Pseudomonas avellanae* species. PLoS One 8:e75794. http://dx.doi.org/10.1371/journal.pone.0075794.
25. Tarazona E, Lucena T, Arahal DR, Macián MC, Ruvira MA, Pujalte MJ. 2014. Multilocus sequence analysis of putative *Vibrio mediterranei* strains and description of *Vibrio thalassae* sp. nov. Syst. Appl. Microbiol. 37:320–328. http://dx.doi.org/10.1016/j.syapm.2014.05.005.
26. Janda JM, Abbott SL. 2010. The genus *Aeromonas*: taxonomy, pathogenicity, and infection. Clin. Microbiol. Rev. 23:35–73. http://dx.doi.org/10.1128/CMR.00039-09.
27. Cheesman SE, Neal JT, Mittge E, Seredick BM, Guillemin K. 2011. Epithelial cell proliferation in the developing zebrafish intestine is regulated by the Wnt pathway and microbial signaling via Myd88. Proc. Natl. Acad. Sci. U. S. A. 108:4570–4577. http://dx.doi.org/10.1073/pnas.1000072107.
28. Martin-Carnahan A, Joseph SW. 2005. Genus I. *Aeromonas Stanier* 1943: 213AL, p 5574570–578. *In* Brenner DJ, Krieg NR, Staley JT, Garrity GM (ed), Bergey's manual of systematic bacteriology, vol 2, 2nd ed. Springer Verlag, New York, NY.
29. Huys G, Cnockaert M, Swings J. 2005. *Aeromonas culicicola* Pidiyar et al. 2002 is a later subjective synonym of *Aeromonas veronii* Hickman-Brenner et al. 1987. Syst. Appl. Microbiol. 28:604–609. http://dx.doi.org/10.1016/j.syapm.2005.03.012.
30. Huys G, Kämpfer P, Swings J. 2001. New DNA-DNA hybridization and phenotypic data on the species *Aeromonas ichthiosmia* and *Aeromonas allosaccharophila*: *A. ichthiosmia* Schubert et al. 1990 is a later synonym of *A. veronii* Hickman-Brenner et al. 1987. Syst. Appl. Microbiol. 24:177–182. http://dx.doi.org/10.1078/0723-2020-00038.
31. Collins MD, Martinez-Murcia AJ, Cai J. 1993. *Aeromonas enteropelogenes* and *Aeromonas ichthiosmia* are identical to *Aeromonas trota* and *Aeromonas veronii*, respectively, as revealed by small-subunit rRNA sequence analysis. Int. J. Syst. Bacteriol. 43:855–856. http://dx.doi.org/10.1099/00207713-43-4-855.
32. Huys G, Denys R, Swings J. 2002. DNA-DNA reassociation and phenotypic data indicate synonymy between *Aeromonas enteropelogenes* Schubert et al. 1990 and *Aeromonas trota* Carnahan et al. 1991. Int. J. Syst. Evol. Microbiol. 52:1969–1972. http://dx.doi.org/10.1099/ijs.0.01996-0.
33. Roger F, Marchandin H, Jumas-Bilak E, Kodjo A, colBVH study group, Lamy B. 2012. Multilocus genetics to reconstruct aeromonad evolution. BMC Microbiol. 12:62. http://dx.doi.org/10.1186/1471-2180-12-62.
34. Martinez-Murcia AJ, Monera A, Saavedra MJ, Oncina R, Lopez-Alvarez M, Lara E, Figueras MJ. 2011. Multilocus phylogenetic analysis of the genus *Aeromonas*. Syst. Appl. Microbiol. 34:189–199. http://dx.doi.org/10.1016/j.syapm.2010.11.014.
35. Miñana-Galbis D, Farfán M, Albarral V, Sanglas A, Lorén JG, Fusté MC. 2013. Reclassification of *Aeromonas hydrophila* subspecies *anaerogenes*. Syst. Appl. Microbiol. 36:306–308. http://dx.doi.org/10.1016/j.syapm.2013.04.006.
36. Huys G, Kämpfer P, Albert MJ, Kühn I, Denys R, Swings J. 2002. *Aeromonas hydrophila* subsp. *dhakensis* subsp. nov., isolated from children with diarrhoea in Bangladesh, and extended description of *Aeromonas hydrophila* subsp. *hydrophila* (Chester 1901) Stanier 1943 (approved lists 1980). Int. J. Syst. Evol. Microbiol. 52:705–712. http://dx.doi.org/10.1099/ijs.0.01844-0.
37. Figueras MJ, Beaz-Hidalgo R, Senderovich Y, Laviad S, Halpern M. 2011. Re-identification of *Aeromonas* isolates from chironomid egg masses as the potential pathogenic bacteria *Aeromonas aquariorum*. Environ. Microbiol. Rep. 3:239–244. http://dx.doi.org/10.1111/j.1758-2229.2010.00216.x.
38. Beaz-Hidalgo R, Martínez-Murcia A, Figueras MJ. 2014. Corrigendum to "Reclassification of Aeromonas hydrophila subsp. dhakensis Huys et al. 2002 and *Aeromonas aquariorum* Martínez-Murcia et al. 2008 as *Aeromonas dhakensis* sp. nov. comb nov. and emendation of the species *Aeromonas hydrophila*." [Syst. Appl. Microbiol. 36:171–176, 2013.] Syst. Appl. Microbiol. 37:543. http://dx.doi.org/10.1016/j.syapm.2012.12.007.
39. Martino ME, Fasolato L, Montemurro F, Rosteghin M, Manfrin A, Patarnello T, Novelli E, Cardazzo B. 2011. Definition of microbial diversity in *Aeromonas* strains based on multilocus sequence typing, phenotype and presence of putative genes of virulence. Appl. Environ. Microbiol. 77:4986–5000. http://dx.doi.org/10.1128/AEM.00708-11.
40. Loren JG, Farfan M, Fuste MC. 2014. Molecular phylogenetics and temporal diversification in the genus *Aeromonas* based on the sequences of five housekeeping genes. PLoS One 9:e88805. http://dx.doi.org/10.1371/journal.pone.0088805.
41. Silver AC, Williams D, Faucher J, Horneman AJ, Gogarten JP, Graf J. 2011. Complex evolutionary history of the *Aeromonas veronii* group revealed by host interaction and DNA sequence data. PLoS One 6:e16751. http://dx.doi.org/10.1371/journal.pone.0016751.
42. Morandi A, Zhaxybayeva O, Gogarten JP, Graf J. 2005. Evolutionary and diagnostic implications of intragenomic heterogeneity in the 16S rRNA gene in *Aeromonas* strains. J. Bacteriol. 187:6561–6564. http://dx.doi.org/10.1128/JB.187.18.6561-6564.2005.
43. Roger F, Lamy B, Jumas-Bilak E, Kodjo A, colBVH Study Group, Marchandin H. 2012. Ribosomal multi-operon diversity: an original perspective on the genus *Aeromonas*. PLoS One 7:e46268. http://dx.doi.org/10.1371/journal.pone.0046268.
44. Beaz-Hidalgo R, Martínez-Murcia A, Figueras MJ. 2013. Reclassification of *Aeromonas hydrophila* subsp. *dhakensis* Huys et al. 2002 and *Aeromonas aquariorum* Martinez-Murcia et al. 2008 as *Aeromonas dhakensis* sp. nov. comb. nov. and emendation of the species *Aeromonas hydrophila*. Syst. Appl. Microbiol. 36:171–176. http://dx.doi.org/10.1016/j.syapm.2012.12.007.
45. Grim CJ, Kozlova EV, Ponnusamy D, Fitts EC, Sha J, Kirtley ML, van Lier CJ, Tiner BL, Erova TE, Joseph SJ, Read TD, Shak JR, Joseph SW, Singletary E, Felland T, Baze WB, Horneman AJ, Chopra AK. 2014. Functional genomic characterization of virulence factors from necrotizing fasciitis-causing strains of *Aeromonas hydrophila*. Appl. Environ. Microbiol. 80:4162–4183. http://dx.doi.org/10.1128/AEM.00486-14.
46. Carnahan AM, Behram S, Joseph SW. 1991. Aerokey II: a flexible key for identifying clinical *Aeromonas* species. J. Clin. Microbiol. 29:2843–2849.
47. Beaz-Hidalgo R, Alperi A, Buján N, Romalde JL, Figueras MJ. 2010. Comparison of phenotypical and genetic identification of *Aeromonas* strains isolated from diseased fish. Syst. Appl. Microbiol. 33:149–153. http://dx.doi.org/10.1016/j.syapm.2010.02.002.
48. Martínez-Murcia A, Monera A, Alperi A, Figueras MJ, Saavedra MJ. 2009. Phylogenetic evidence suggests that strains of *Aeromonas hydrophila* subsp. *dhakensis* belong to the species *Aeromonas aquariorum* sp. nov.

Curr. Microbiol. 58:76–80. http://dx.doi.org/10.1007/s00284-008-9278-6.
49. Giltner CL, Bobenchik AM, Uslan DZ, Deville JG, Humphries RM. 2013. Ciprofloxacin-resistant *Aeromonas hydrophila* cellulitis following leech therapy. J. Clin. Microbiol. 51:1324–1326. http://dx.doi.org/10.1128/JCM.03217-12.
50. Martínez-Murcia A, Beaz-Hidalgo R, Svec P, Saavedra MJ, Figueras MJ, Sedlacek I. 2013. *Aeromonas cavernicola* sp. nov., isolated from fresh water of a brook in a cavern. Curr. Microbiol. 66:197–204. http://dx.doi.org/10.1007/s00284-012-0253-x.
51. Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. Proc. Natl. Acad. Sci. U. S. A. 102:14332–14337. http://dx.doi.org/10.1073/pnas.0504068102.
52. Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492–508. http://dx.doi.org/10.1080/10635150290069913.
53. Andam CP, David W, Gogarten JP. 2010. Biased gene transfer mimics patterns created through shared ancestry. Proc. Natl. Acad. Sci. U. S. A. 107:10679–10684. http://dx.doi.org/10.1073/pnas.1001418107.
54. Pace NR, Sapp J, Goldenfeld N. 2012. Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. Proc. Natl. Acad. Sci. 109:1011–1018. http://dx.doi.org/10.1073/pnas.1109716109.
55. Williams D, Fournier GP, Lapierre P, Swithers KS, Green AG, Andam CP, Gogarten JP. 2011. A rooted net of life. Biol. Direct 6:45. http://dx.doi.org/10.1186/1745-6150-6-45.
56. Bansal MS, Alm EJ, Kellis M. 2013. Reconciliation revisited: handling multiple optima when reconciling with duplication, transfer, and loss. J. Comput. Biol. 20:738–754.
57. García JA, Larsen JL, Dalsgaard I, Pedersen K. 2000. Pulsed-field gel electrophoresis analyis of *Aeromonas salmonicida* ssp. *salmonicida*. FEMS Microbiol. Lett. 190:163–166. http://dx.doi.org/10.1111/j.1574-6968.2000.tb09280.x.
58. Thompson CC, Vicente ACP, Souza RC, Vasconcelos ATR, Vesth T, Alves N, Jr, Ussery DW, Iida T, Thompson FL. 2009. Genomic taxonomy of vibrios. BMC Evol. Biol. 9:258. http://dx.doi.org/10.1186/1471-2148-9-258.
59. Chen PL, Wu CJ, Chen CS, Tsai PJ, Tang HJ, Ko WC. 2014. A comparative study of clinical *Aeromonas dhakensis* and *Aeromonas hydrophila* isolates in southern Taiwan: *A. dhakensis* is more predominant and virulent. Clin. Microbiol. Infect. 20:O428–O434. http://dx.doi.org/10.1111/1469-0691.12456.
60. Sambrook J, Russell DW. 2001. Molecular cloning: a laboratory manual, 3rd ed. Cold Spring Harbor, New York, NY.
61. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R. 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 42:D206–D214. http://dx.doi.org/10.1093/nar/gkt1226.
62. Bomar L, Stephens WZ, Nelson MC, Velle K, Guillemin K, Graf J. 2013. Draft genome sequence of *Aeromonas veronii* Hm21, a symbiotic isolate from the medicinal leech digestive tract. Genome Announc. 1(5):e00800-13. http://dx.doi.org/10.1128/genomeA.00800-13.
63. Gouy M, Guindon S, Gascuel O. 2010. SeaView, version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27:221–224. http://dx.doi.org/10.1093/molbev/msp259.
64. Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113. http://dx.doi.org/10.1186/1471-2105-5-113.
65. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML. Syst. Biol. 59:307–321. http://dx.doi.org/10.1093/sysbio/syq010.
66. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Paley S, Popescu L, Pujar A, Shearer AG, Zhang P, Karp PD. 2010. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 38:D473–D479. http://dx.doi.org/10.1093/nar/gkp875.
67. van Dongen S, Abreu-Goodger C. 2012. Using MCL to extract clusters from networks. Methods Mol. Biol. 804:281–295. http://dx.doi.org/10.1007/978-1-61779-361-5_15.
68. Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30:1575–1584. http://dx.doi.org/10.1093/nar/30.7.1575.
69. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. http://dx.doi.org/10.1371/journal.pone.0009490.
70. Swofford DL. 2002. PAUP*: phylogenetic analysis using parsimony (and other methods), 4th ed. Sinauer Associates, Sunderland, MA.
71. Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology open software suite. Trends Genet. 16:276–277. http://dx.doi.org/10.1016/S0168-9525(00)02024-2.
72. Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690. http://dx.doi.org/10.1093/bioinformatics/btl446.
73. Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247. http://dx.doi.org/10.1093/bioinformatics/17.12.1246.
74. Seshadri R, Joseph SW, Chopra AK, Sha J, Shaw J, Graf J, Haft D, Wu M, Ren Q, Rosovitz MJ, Madupu R, Tallon L, Kim M, Jin S, Vuong H, Stine OC, Ali A, Horneman AJ, Heidelberg JF. 2006. Genome sequence of *Aeromonas hydrophila* ATCC 7966^T: jack of all trades. J. Bacteriol. 188:8272–8282. http://dx.doi.org/10.1128/JN.00621-06.
75. Spataro N, Farfán M, Albarral V, Sanglas A, Lorén JG, Fusté MC, Bosch E. 2013. Draft genome sequence of *Aeromonas molluscorum* strain 848TT, isolated from bivalve molluscs. Genome Announc. 1(3):e00382-13. http://dx.doi.org/10.1128/genomeA.00382-13.
76. Beatson SA, das Graças de Luna M, Bachmann NL, Alikhan NF, Hanks KR, Sullivan MJ, Wee BA, Freitas-Almeida AC, Dos Santos PA, de Melo JT, Squire DJ, Cunningham AF, Fitzgerald JR, Henderson IR. 2011. Genome sequence of the emerging pathogen *Aeromonas caviae*. J. Bacteriol. 193:1286–1287. http://dx.doi.org/10.1128/JB.01337-10.
77. Wu CJ, Wang HC, Chen CS, Shu HY, Kao AW, Chen PL, Ko WC. 2012. Genome sequence of a novel human pathogen, *Aeromonas aquariorum*. J. Bacteriol. 194:4114–4115. http://dx.doi.org/10.1128/JB.00621-12.
78. Chan KG, Puthucheary SD, Chan XY, Yin WF, Wong CS, Too WS, Chua KH. 2011. Quorum sensing in *Aeromonas* species isolated from patients in Malaysia. Curr. Microbiol. 62:167–172. http://dx.doi.org/10.1007/s00284-010-9689-z.
79. Tekedar HC, Waldbeiser GC, Karsi A, Liles MR, Griffin MJ, Vamenta S, Sonstegard T, Hossain M, Schroeder SG, Khoo L, Lawrence ML. 2013. Complete genome sequence of a channel catfish epidemic isolate, *Aeromonas hydrophila* strain ML09-119. Genome Announc. 1(5):e00755-13. http://dx.doi.org/10.1128/genomeA.00755-13.
80. Han JE, Kim JH, Choresca C, Shin SP, Jun JW, Park SC. 2013. Draft genome sequence of a clinical isolate, *Aeromonas hydrophila* SNUFPC-A8, from a moribund cherry salmon (*Oncorhynchus masou masou*). Genome Announc. 1(1):e00133-12. http://dx.doi.org/10.1128/genomeA.00133-12.
81. Chai B, Wang H, Chen X. 2012. Draft genome sequence of high-melanin-yielding *Aeromonas media* strain WS. J. Bacteriol. 194:6693–6694. http://dx.doi.org/10.1128/JB.01807-12.
82. Han JE, Kim JH, Shin SP, Jun JW, Chai JY, Park SC. 2013. Draft genome sequence of *Aeromonas salmonicida* subsp. *achromogenes* AS03, an atypical strain isolated from crucian carp (*Carassius carassius*) in the Republic of Korea. Genome Announc. 1:e00791-13. http://dx.doi.org/10.1128/genomeA.00791-132229.
83. Reith ME, Singh RK, Curtis B, Boyd JM, Bouevitch A, Kimball J, Munholland J, Murphy C, Sarty D, Williams J, Nash JH, Johnson SC, Brown LL. 2008. The genome of *Aeromonas salmonicida* subsp. *salmonicida* A449: insights into the evolution of a fish pathogen. BMC Genomics 9:427. http://dx.doi.org/10.1186/1471-2164-9-427.
84. Charette SJ, Brochu F, Boyle B, Filion G, Tanaka KH, Derome N. 2012. Draft genome sequence of the virulent strain 01-B526 of the fish pathogen Aeromonas salmonicida. J. Bacteriol. 194:722–723. http://dx.doi.org/10.1128/JB.06276-11.
85. Li Y, Liu Y, Zhou Z, Huang H, Ren Y, Zhang Y, Li G, Zhou Z, Wang L. 2011. Complete genome sequence of *Aeromonas veronii* strain B565. J. Bacteriol. 193:3389–3390. http://dx.doi.org/10.1128/JB.00347-11.

Chapter 3 – Extension of the ANI concept to generate phylogenies

This section consists of a not yet submitted draft manuscript detailing a project Sean Gosselin and I have been pursuing. The central goal is to determine if average nucleotide identity data can be used to infer whole-genome phylogenies of a similar quality to more complex methodologies. To this effect we developed metric that extends the ANI concept as well as a bootstrapping method, which provides statistical support for our phylogeny inference. As a carry-on effect we found our metric provided useful and informative cut-offs for delimiting genera and family-level taxonomic ranks.

This project was conceived by myself and J. Peter Gogarten. Sean and I planned and executed the analyses. Sean was primarily responsible for the coding of the ANI-methodology and was behind several of the statistical analyses. He also made most of the qualitative tree comparisons. He participated in writing and editing of the manuscript. I coded the *is*DDH-methodology, performed almost all of the tree-building, and tree-related statistical analysis. I composed and wrote the first draft of the manuscript and have continued to write and edit it. I also wish to acknowledge and thank Yutian Feng for contributing assistance with the *Frankiales* MLSA phylogeny.

Chapter 3.1 Expanding the Utility of Comparisons Using Data From Whole Genomes

Expanding the Utility of Comparisons Using Data From Whole Genomes

Matthew S. Fullmer[1]*, Sean Gosselin[1]*, and J. Peter Gogarten[1,2]

[1] Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA

[2] Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA

* Authors contributed equally

Abstract

Whole genome comparisons, Average Nucleotide Identities (ANI) and the Genome-to-genome distance calculator (isDDH) have risen to prominence in rapidly classifying taxa using whole genome sequences. The jSpecies implementation of ANI in particular, has been proposed to be a new standard in species classification, and has become a common technique for papers including newly sequenced genomes. However, attempts to use whole genome comparison data to infer phylogenies have had difficulty matching those produced by more complex phylogenetics methods. We present two novel methods for generating reliable and statistically supported phylogenies using ANI and isDDH data matching established techniques. The isDDH method returns good results up to approximately the genus level while the ANI method extends to at least the family level. These two novel methods offer the opportunity make use of whole-genome comparison data that is already being generated to produce relatively quick and accurate phylogenies. As a final bonus, the developed ANI methodology also offers the ability to

delimit cut-offs for deeper taxonomic ranks than the species level ANI and isDDH is usually confined to.

Introduction

DNA-DNA Hybridization (DDH) holds the distinction of being the gold standard of gold standards for species delineation (Stackebrandt and Goebel, 1994). The method is technically challenging and its results are frequently poorly reproducible across labs. As a result, there have been ongoing efforts to supplement or replace DDH with *in silico* methods taking advantage of the ongoing revolution in genome sequencing (Auch et al., 2010a; Goris et al., 2007; Konstantinidis and Tiedje, 2005; Varghese et al., 2015). The two major approaches have been Average Nucleotide Identity (ANI) and the Genome-to-Genome Distance Calculator (GGDC) (Auch et al., 2010a; Konstantinidis and Tiedje, 2005).

From its beginning the proportion of a genome in common with another has been a major element in ANI thinking. When first proposed in 2005 the method used the average identity of shared ORFs (Konstantinidis and Tiedje, 2005). After defining a pure ANI cutoff the authors then examined the large disparities in gene content among the strains and species in their dataset. A year later they explored the question in much greater depth and observed that the ANI was correlated with the percent of content shared (Konstantinidis et al., 2006). i.e., that it seemed a significant amount of divergence needed to have occurred before major shifts in genome content occurred. In 2007, the emphasis was shifted from the ORFs to the whole genome as the ANI method was adapted to directly compare to DDH (Goris et al., 2007). This whole-genome approach became a popular method of performing the technique. For example, the jSpecies Java application was developed to perform the Goris method in a local and scalable manner

(Richter and Rosselló-Móra, 2009). However, the consideration of the varying gene content became de-emphasized with the default exportable output from jSpecies not including any reference to shared content in a comparison (although, the percent used in the comparison was available for viewing in the GUI interface and in the raw data files). The de-emphasis on gene content is largely irrelevant when comparing closely related organisms. As noted above, there is a correlation between ANI and shared genome content. ANI results can give spurious and misleading results when only small fractions of the genomes are shared. The gene content issue did return in 2015 with the publication of the gANI method (Varghese et al., 2015). This approach explicitly considers the shared content and offers two separate delimiters for a species: ANI (as calculated from ORFs, but also considering the comparative length of each side in the bidirectional best hits.) as well as an "Alignment Fraction" or proportion of genes shared. While gANI offers an important upgrade to the ANI paradigm it does contain an important limitation. Namely, how does one interpret a comparison between two taxa where the ANI is above the threshold and the AF is below? And vice-versa? One wonders if these two measures could be combined into a single metric that encompasses both.

Generating trees from whole genome distances has been common for quite some time (Gibbon et al., 1999). However, they have often suffered from a lack of statistical support for their branching patterns (Krajewski and Dickerman, 1990). It has been common to use a set of MLSA-style genes or a strict/relaxed core gene set to build a phylogeny that proxies for the signal of the whole genome; for example, the seminal papers in the development of ANI (Goris et al., 2007; Konstantinidis and Tiedje, 2005). MLSA

methods, while generally strong, have weaknesses. Firstly, the choice of genes is important. One needs an a priori set of single copy informative genes. These genes are assumed to be rarely horizontally transferred by the complexity hypothesis (Jain et al., 1999) as well as representing broadly the same history. Unfortunately, more conserved genes are often more frequently transferred among closely related organisms (Gogarten et al., 2002). Likewise, the genes do not always agree wholeheartedly on their histories (Colston et al., 2014; Fullmer et al., 2014; Salichos et al., 2014). Core and relaxed core genomes offer advantages by incorporating a tremendous amount of overall information. Their primary drawback is the need to annotate the whole genome (albeit a fairly modest tribulation), cluster the genes into homologous groups (the largest obstacle, especially with large numbers of genomes), and finally the computational power required to compute a supported phylogeny from such a large amount of data.

Alongside the evolution of ANI, with origins actually predating it (Henz et al., 2005), another method was developing. This method, realized in the Genome-to-Genome Distance Calculator (GGDC), provides whole genome measures directly on the same scale as DNA-DNA Hybridization (DDH). This is in contrast to ANI methods that often fail to correlate linearly with observed DDH values. I.e., ANIs translated to DDH scale can rise above 100% and below zero percent DDH, the maximum and minimum possible experimental DDHs. A prime advantage of the GGDC methods is that they also calculate a 95% confidence interval. This provides an inbuilt measure of support to the point estimates that ANI measures lack. GGDC also provides results from 3 formulas aimed at different levels of sequencing completeness, allowing some opportunity to provide more support and assessment of the output.

We postulated that we might be able to generate supported phylogenies from whole genome comparison data. We aimed to create a single distance measure from ANI data incorporating both the ANI and genome proportion. A single measure simplifies treebuilding as well as interpretation of the dissimilarity of the genomes. We label this new ANI measure Total Average Nucleotide Identity (tANI). We then developed resampling methods to create bootstrapped sets of raw data. Using these methods we gain the advantages of a core genome phylogeny while skipping the onerous annotation and gene clustering processes. We would also gain process time back by being able to use distance-based tree-building methods that are typically far less time consuming than maximum-likelihood or Bayesian inference methods.

Result

Checking Parameters

Bootstrap confidence sets are reliable for tANI

Our treebuilding results demonstrate that our methods can produce phylogenies that hold up well to traditional methods such as MLSA at genus levels while tANI can perform well up to approximately the order or class level. However, one concern is whether our method's bootstrap values are similar to the traditional methods. To assess the uncertainty of the support sets Internode Certainty (IC) scores were calculated by mapping support sets against reference trees as implemented in RAxML v8.1 (Salichos et al., 2014; Stamatakis, 2014). IC represents a quantification of the level of disagreement in a support set for a particular node in a phylogeny. The IC decreases as the bootstrap value drops and as the dissenting samples agree on fewer rival topologies. I.e., a bipartition supported by 51% of the bootstraps would score much higher if the conflicting samples represented 49 different alternatives than it would if there were only a single alternative. The tree certainty average value (TCA) is the average of IC values across the entire tree, representing an assessment of overall conflict in the support set. The mBio dataset was used as a test case as it already offers an expanded core phylogeny in addition to the MLSA, allowing a more appropriate comparison between whole genome methods. The TCA for the mBio MLSA was 0.65 and 0.86 for the tANI phylogeny confidence set. This suggests that these two datasets are capturing a dissimilar amount of uncertainty. However, the bootstrap set for the expanded core genome from the mBio dataset shows a TCA of 0.87, very similar to the tANI data. These data are capturing a very similar amount of uncertainty, possibly much of the same uncertainty. When the confidence sets

are paired with the other method's best tree the TCAs are 0.61 for both combinations, suggesting much of the uncertainty being observed is shared between the two datasets. To take another look at the similarity, or not, between the expanded core and the ANI data the distances between the topologies of the two confidence sets were compared and projected via PCoA. This projection **(Figure01)**, shows that the clusters of one dataset collocate with the other. The upshot is the suggestion that our novel treebuilding method is capturing a very comparable amount of uncertainty in the whole genome data as that observed in the established methodology demonstrated by the mBio data.

The Internode Certainty (IC) score implemented in the RAxML package (Salichos et al., 2014; Stamatakis, 2014) affords a measure of this similarity. IC represents a quantification of the level of disagreement in a support set for a particular node in a phylogeny. Mapping the bootstrap support set from our tANI methodology onto the mBio core genome phylogeny, and vice-versa, revealed average IC (ICA) scores of 0.61. This is similar to the ICA score when the mBio MLSA support set is mapped back upon itself. It is also relatively close to the core support set mapped onto itself (0.87). This suggests that the tANI and core genome methods are capturing a similar level of uncertainty in the data, and largely the same uncertainty.

Genomic Characteristics Do Not Bias Results

These methods utilize whole genome assemblies in their calculation of distances. Differences in genomic traits such as size and GC-content could conceivably bias the

results of the calculations and introduce artefacts into the final phylogenies and support values. To test these possibilities we developed a fourth test set using the order Frankiales, composed primarily of the genus *Frankia*. This group was chosen on account of their extreme variance in genome size (~4Mb to ~11 Mb) and considerable range of GC-contents (~60% to ~75%).

The *Frankia* set did not produce a radically different tree from our MLSA-derived reference phylogeny (**FigureS01)**. This suggests that we will be able to make fair use of it for examining the genome size and GC-content question. When comparing the tANI phylogeny against genome size (**Figure02**), there is no pattern of clustering by genome size. Some groups cluster with similar sizes, such as the *F. coriariae* and *F. alni* clades. However, these match the MLSA topology and also fail to either attract to other groups with similar genome size, or repulse from groups that are different. When examining the GC-content, we see a very similar result (**Figure03**). There are no obvious patterns of GC-content biasing taxa together at the expense of their presumed placement.

Saturation Limits jANI Compared to Our Method

As jANI (Richter and Rosselló-Móra, 2009) has become a standard in many studies, we compare our method against it. jANI values in the absence of shared genomic content can lead to erroneous conclusions. There are two causes. First, as the distance between two genomes increases the fractions of the genome included in the ANI calculation drops rapidly. This is often not a problem among the most closely related of organisms as shared content and ANI correlate strongly at high ANIs (Konstantinidis and Tiedje, 2005). Second, as genomic divergence increases the impact of sequence

saturation increases, decreasing the apparent distances. As genome comparisons move away from the species-to-species scale that NI was designed for, the noise in the jANI result can become considerable (**FigureS02**). At extreme levels, the jANI values can border on species cutoffs despite incorporating only a miniscule fraction of the genomes (~0.1%). An example of this occurs in the Roseo dataset. Rhodobacterales bacterium HTCC 2255 demonstrates jANI values as high as 94%, average of 89%, and a median score of 91% (the average and median in that dataset are both 83%). This effect can also be seen the topology produced from a distance tree inferred from uncorrected jANI values (**FigureS03**). The result for the Roseo set clearly shows a topology with multiple differences from Collins et al., demonstrating the effect of saturation on phylogeny beyond the most closely related of taxa.

Our novel tANI method ameliorates these issues by incorporating the alignment fraction into a final distance value and applying a saturation correction. Our value increases while jANI enters the early stages of saturation in the neighborhood of 85% identity. If using jANI with the MUMmer algorithm the saturation effects appear even earlier. The result is that researchers studying new isolates that are not known to be especially close relatives to each other or references may inadvertently arrive at incorrect conclusions.

Accuracy of Novel Methods Against Multi-gene Methods

tANI

Trees calculated by our tANI methodology showed a high degree of agreement with more established methods (**Figure01**). For the mBio set our distance based tree shows a high

degree of convergence in bran4hing pattern with that of the core genome maximum likelihood tree produced in the original manuscript (Colston et al., 2014). Looking at specific cases, there are small differences in the placement of the *A. veronii* AMC34 and the *A. allosaccharophila* clade. The branches leading to this node are poorly supported (**Figure04**), but the deeper clades are maintained and highly supported (>90%). The similarity of the two trees holds up across different cutoff values for filtering the best BLAST hits for the calculation (data not shown).

Our distance-based methods also produce results comparable to traditional methods when examining ambiguous clades. In the trees produced from our Roseobacter set genomes (**Figure05**), we see highly questionable support values for the largest clade (highlighted in blue) from both the MLSA tree and our distance-based tree (30/100,45/100 respectively).

The AeroOG set, which includes all of the publicly available *Aeromonadales* genomes as of January 2017, suggests that the higher order classification of *Aeromonadales* may also be up for debate. Members of the *Succinivibrionaceae* are extremely distant from the rest of our data set, to the point where our distance value begins to become unreliable. For example, these distance values are on par with or higher than *Enterobacteriaceae* when compared to members of the genus *Aeromonas* (**FigureS04**). The individual *Succinivibrionaceae* may be grouping together as the result of long branch attraction, yet it is hard to look at the family in higher detail, as there are few sequences currently available on NCBI. In addition, the original classification of this order did not include the *Succinivibrionaceae* and no further analyses were reported that

confirmed they should be included (Martin-Carnahan and Joseph, 2015; Stackebrandt and Hespell, 2006).

Internode Certainty mapping also provides a degree of quantification to these conclusions. Mapping the mBio tANI bootstraps onto the tANI tree and mBio core genome bootstraps to core genome tree both yield a result of ~0.86 average IC across all nodes. When the ANI support set is mapped against the core tree the score still stands at 0.61, indicating a substantial amount of support for the core tree topology. This stands in contrast to mapping MLSA supports onto the ANI tree, which maintained only a 0.33 score. There is still agreement with the ANI topology, but the conflicting signal is much stronger, fitting with the results reported in Colston et al., 2014. The Roseo set tells a similar, if slightly less ideal picture. Self-to-self mapping return IC values of ~0.80. While, mapping MLSA bootstraps onto the ANI topology produces a reduced IC average score of 0.39. While this clearly indicates the two methods are exploring at least a substantially similar area of tree-space, they are less in line than the mBio dataset.

isDDH

The isDDH treebuilding on the mBio set produced results largely in concordance with Colston et al., 2014, but with a greater number of topological differences supported by strong bootstrap values (**Figure06**). For example, the nearest neighbor of *A. veronii*, is *A.* sp. AMC34 at 87% bootstraps. Additionally, *A. sobria* falls out near the base of the so-called Aeromonas veronii group rather than grouped with the two *A. allosacharophila* strains.

Inferring a phylogeny for the Roseo set, the results showed less in common with Collins et al. (**Figure07**). While groups such as the *Leisingera* and *Phaeobacter* fell out

as would be expected a number of others, including the *Ruegeria* and *Tateymaria* not matching the published topology. More concerning was that a number of nodes received negligible or even zero support from the confidence set, suggesting that the resampling and treebuilding method has reached or passed the limits of its usefulness.

Treebuilding with the isDDH method at the higher taxonomic levels, with the AeroOG set, confirmed the results of the Roseo set. Placements of taxa border on random in many cases. For example, *A. lacus* resides on one of the branches furthest from the center of *Aeromonas* gravity. *V. cholerae* IEC224 and *Succinatimonas hippei* YIT12066 lie within the *Aeromonas* genus. Most of the accepted genera are split up in multiple locations across the tree. The only reasonable conclusion is that our method for inferring phylogenies from isDDH data lacks the potential of our ANI method. This is certainly true as one moves into taxonomic ranks deeper than genus level.

Misclassified Taxa

A number of taxa in our datasets appear to be misclassified (**TableS01**). Almost all of these taxa fall into groups with which there exists phylogenetic support for their misclassification. These taxa were reclassified into novel groups along their phylogenetic lines for the purpose of our taxonomic rank cut-off analyses. Our tANI metric agreed with these decisions and its sensitivities and specificities improved as a result.

This Novel Extension of ANI Matches Older Methodologies

To determine the cutoff for a species based on a single genome to genome distance calculation we used a receiver operating characteristic curve (ROC) analysis. Working on

the union of the mBio and Roseo the ROC estimates a distance cutoff of .4422320, at a specificity of 99.70813, and sensitivity of 100.00 (**Figure08a**) against the accepted nomenclature. Examination of the ROCs for the constituent datasets reveals that the two genera are not equally easy to classify (**Figure08bc**). However, when taxa in the Roseo set are reclassified along the lines illustrated in the MLSA phylogeny the genus's curve responds favorably, improving the sensitivity from 80% to 99% (**Figure 08d**).

Our Novel tANI Method Offers the Ability to Delimit Deeper Taxonomic Ranks

One side effect from our use of broader taxonomic samplings in some of our datasets is the opportunity to test our distance measure against those units *sensu* ANI and GGDC species cutoffs. When we plotted the distances for every interaction with the union of the AeroOG and Roseo sets we observed a series of recognizable peaks for each taxonomic rank (**Figure09**). Particularly, once the taxa suspected of misclassification were reclassified. We made use of the ROCs once again to provide statistical evidence for these observations. At the genus level, the AeroOG set (**Figure10a)** and Roseo set (**Figure10b**) have similar, although not identical distance cutoffs (3.28 and 3.25, respectively) and varied but generally high specificities (96.7% and 86.8%) and sensitivities (99.5% and 95.4%). Zooming out to family distinctions, the combined datasets returned a cutoff of 4.57, and maintained specificity of 90.7% and sensitivity of 86.7% (**Figure10c**), suggesting an ability to still discriminate relationships at this level.. At order level **(Figure10d)** the combined datasets fell off to 4.48 and 93.5% and 78.9%, suggesting the method no longer could reliably discriminate at this taxonomic rank.

Comparison of tANI Method With Other Whole-Genome Methods

Building trees from whole-genome data is hardly a novel concept. As such, it is important for us to compare our methodology with some of the other methods available and assess our methodologies strengths and weaknesses. We chose two approaches. First is the Genome BLAST Distance Phylogeny (GBDP) (Auch et al., 2006; Henz et al., 2005). This method is highly relevant as the *is*DDH method we tested uses the GGDC 2.1 freely available on the DSMZ website (http://ggdc.dsmz.de/home.php). This service runs on an updated version of the GBDP making it in some respects a spiritual sister to our *is*DDH concept. Also relevant, the updated GBDP software (Meier-Kolthoff et al., 2014) has recently been used as part of several recent large studies (Hahnke et al., 2016; Mukherjee et al., 2017; Peeters et al., 2016) showing it to be a present player in the whole-genome phylogenetics field. The second method is Mashtree (https://github.com/lskatz/mashtree), which is an extension of the Mash kmer-calculation software (Ondov et al., 2016).

Unfortunately, the only stand-alone version of GDBP we are aware of is a legacy beta version (http://www.auch-edv.de/GBDP). The GGDC website reports that an improved standalone will be available sometime in 2018. But for our comparison we have had to content ourselves with what was in the available package. We compared the GDBP using the mBio, Roseo, and AeroOG datasets. GBDP produces outputs using a range of possible distance equations. In all cases we chose the method that was most treelike when viewed in SplitsTree v4. For both datasets the topology was far from expectations **(FigureS05, panels A and B)**. In the AeroOG dataset, the GBDP tree failed

to correctly identify the *Aeromonas* as a single clade, separating out major portions closer to the other Gammaproteobacteria. Additionally, GBDP placed large portions of the *Oceanimonas* group in the middle of the largest *Aeromonas* clade. GBDP trees were equally suspect when calculated for the *Roseobacter* dataset. GDBP trees incorrectly grouped the *Leisingera* genus into three separate clades, while also splitting *Loktanella* and *Ruegeria* into countless different clades. However, GBDP does work well when looking at within genera phylogenies. For the MBio dataset, the GBDP based tree largely agreed on the clustering of various clades such as the *hydrophila*/*dhakensis* group, but disagreed on the internal branching between these clades. Not only were the higher taxonomic rank topologies quite different from the ANI and other references but mapping the ANI and reference support sets onto these topologies produced average IC values ranging from indifferent (0.014 with the mBio MLSA support set) to slightly supporting a majority contradictory topology (-0.225 with the Roseo MLSA support set) **(Table01)**.

These results were surprising given the GBDP's usage in a number of studies studies (Hahnke et al., 2016; Mukherjee et al., 2017; Peeters et al., 2016) and its ongoing role at the heart of the fantastic GGDC 2.1 web service. We are left to speculate that the use of a legacy beta version may be to blame and look forward to the release of the updated version later this year to make more meaningful comparisons.

In contrast to the GDBP results, Mashtree performed much more in line with expectations. For the set of AeroOG, MashTree mostly had only small disagreements with our method. For example, MashTree moved the placement of *A. media*, and shuffled members in the *salmonicida*/*aquatica* group. This pattern generally repeats itself in the *Roseobacter* dataset. MashTree successfully kept *Leisingera, Rhodobacter* and the

major *Ruegeria* clade together. Additionally, the MashTree phylogeny generally agrees with the branching patterns our tree proposes, while deviating mostly at nodes of low support. However, MashTree did separate *Loktanella* into a number of monophyletic clades, whereas our method has some of these groupings as polyphyletic. Despite generally matching phylogenies produced by our method, MashTree had significant disagreements when it came to the MBio dataset. Here, MashTree disagreed on internal branching significantly, and shuffled around a number of clades, especially those close to the *A. simiae* clade at the root. When we mapped the mBio tANI support set onto the Mashtree topology the average IC score was 0.472. It scored a 0.283 against the MLSA set, which is comparable to tANI against the same (0.327 ICA) **(Table01)**. Using the Roseo set came up a little less impressive with scores of 0.250 and 0.229, respectively. While this might suggest a weakness for datasets with a deeper taxonomy, the vs. MLSA value was only marginally worse than the tANI-method (0.386).

Overall, GBDP appears to be a little bit "near-sighted" and MashTree a little bit "far-sighted." That is, GBDP appeared to struggle more as the divergence of the dataset increased, while MashTree performed well at larger divergences and came into increasing inaccuracy as the phylogenetic scale shrank.

Methods

Genomes used. Genomes used are listed in the genomes tables (**TableS02)**. Selection initially centered on two groups for which previous phylogenetic and phylogenomic work had been done by this group. The first, hereafter referred to as "mBio" encompasses the 56 *Aeromonas* genomes used in Colston et al., 2014 and represents a genus level taxonomic unit. The second, "Roseo," encompasses those used in Collins et al., 2015 and Gromek et al., 2016 plus additions to investigate the cases of *Loktanella* and *Ruegeria*. This set corresponds closely to a family level taxonomic unit (exempting the genera: *Phenylobacterium, Parvularcula, Maricaulis, Hyphomonas, Hirschia, Caulobacter, Brevundimonas,* and *Asticcacaulis*) A third set, aimed at encompassing a broader phylogenetic and taxonomic range was created by adding all publically available non-*Aeromonas* Aeromonadales genomes to a subset of the mBio set, called the AeroOG set. As the name implies, this set corresponds to an order level unit. Finally, the available genomes from the order *Frankiales* were formed into another dataset with the intention to test the robustness of the ANI method to heterogeneous genome sizes and GC-contents, called the *Frankia* set.

The genomes used in this study are either draft whole genome assemblies or complete assemblies available via NCBI. Those genomes originally sequenced or assembled locally were not systematically screened for plasmids. As doing so would be unlikely to reliably identify all present in our data, when NCBI-derived genomes contained plasmids these were retained with their parent genome. In this way, all of the input data is treated identically. This avoids possible biases introduced by including only some plasmid-equipped taxa while risking the possible bias from allowing plasmids to contribute to our

calculations.

Reference Phylogenies

Comparison reference phylogenies were obtained or generated for each dataset. For mBio, the MLSA and expanded core phylogenies were obtained from Colston et al., 2014. A reference for the Roseo dataset was generated by replicating the method described in Collins et al., 2015, but with added *Loktanella* and *Ruegeria* genomes from NCBI. The AeroOG received its comparator by following the MLSA methodology described in Colston et al., 2014 for the included genomes.

Finally, the Frankia reference required the *de novo* creation of an MLSA scheme in the absence of thorough examples in the literature. 24 single-copy housekeeping genes were selected to form the scheme (**TableS03**). Nucleotide sequences for each gene were retrieved via BLAST, from *Frankia casuarinae* (NC_007777.1). Whole nucleotide sequences for all *Frankia* genomes in genbank format. The program blastn (BLASTALL Version 2.6.0) (Altschul et al., 1997) was executed with the gene sequences as the query and the genomes as the target sequence. The coding sequences corresponding to highest scoring hits for each gene in a singular genome were aligned and concatenated. This was repeated for every genome, generating the multi-locus sequence alignment (MLSA) file. IQTree (Version 1.5.5) was executed with the MLSA file and built the phylogenetic tree with 1000 ultrafast-bootstraps (Chernomor et al., 2016; Haeseler et al., 2017; Hoang et al.; Nguyen et al., 2015). Modeltest arrived at the SCHN05 empirical codon model with empirical codon frequencies (+F) and Free Rate (Soubrier et al., 2012) model of rate heterogeneity with nine categories (+R5).

Description of isDDH support set generation and treebuilding method

Raw data. Estimates of *in silico* DDH were obtained from the GGDC 2.1 program (Auch et al., 2010a, 2010b; Meier-Kolthoff et al., 2013)hosted on the DSMZ website (http://ggdc.dsmz.de/distcalc2.php). Uploads were assembled contigs sets for draft genomes and full chromosome/plasmid sets for completed genomes (**see TableS02**).

Conversion of *is*DDH to Distances. *is*DDH values were converted to distances via the formula: divo = (1 - (DDH value / 100)). A saturation correction was then applied such that the final value: divc = -log(1-divo/max). Where max equals 1.

Creating Bootstrapped Values. Bootstrapped distance matrices were generated through utilization of the 95% confidence interval provided by the GGDC output. The CI provided by the GGDC 2.1 is no longer an explicitly normal distribution. Nonetheless, the upper and lower bounds remain close to symmetrical around the point estimate with a mean difference of only 0.104% (average distances from the point estimates were: 2.41, lower bound and 2.50 for the upper bound) in the *Aeromonas* dataset. A series of highly negatively skewed-normal distributions were tested on the *Aeromonas* data, showing no impact on the topology support values (data not shown). Thus, we consider an assumption of a standard distribution to represent 95% CI to be a reasonable approximation. The shape of this distribution is estimated by extrapolating a standard deviation from the 95% CI. The mean (*is*DDH point estimate) and sigma are then used as parameters for the R function *rnorm* (R Development Core Team, 2008). This function

generates a random value from a normal distribution described by its input parameters. The resulting value is then a resampled *is*DDH value in what is effectively a bootstrapped matrix. The values in these matrices were converted into distances as described above. This procedure is carried out at every position in the matrix such that every genome comparison has been resampled and until the desired number of bootstrapped matrices has been created.

*Specifically, SD = ((CI range * 1.004736564) / 4) The 95% CI is a range slightly smaller than +-2 standard deviations, when assuming a normal distribution. By dividing the 95% CI by 4, the value of a standard deviation may be approximated. Because of how GGDC 2.1 calculates its 95% CI, there is no population size to use to back-calculate the SD, so this approximation is used in its stead.

Description of ANI-extension, support set generation and treebuilding method

ANI and AF Calculation. ANI is calculated in a similar methodology to that described by Varghese et al. (2015) such that ANI is not simply the sum of best hit identities over the total number of genes, but is instead described by the formula:

$$ANI = \frac{\sum(ID\% * Length\ of\ Alignment)}{\sum(Length\ of\ the\ Shorter\ Segment\ from\ BLAST\ Hit)}$$

Alignment fraction is described as:

$$AF = \frac{\sum(Length\ of\ the\ Shorter\ Segment\ from\ BLAST\ Hit)}{Total\ Length\ of\ the\ Query\ Genome}$$

Our methodology differs from Varghese et al. in two respects. First, we do not limit our search to open reading frames but rather use the full scaffold/contig set of an organism. Second, we also fracture the genomes into 1020 nt fragments, in line with previous iterations of ANI calculation (Konstantinidis and Tiedje, 2005; Richter and Rosselló-

Móra, 2009). The fragments from the query genome were each compared to the reference genome via BLAST. After BLAST was completed distance matrices were calculated. Results were filtered based on coverage and percent identity values, and then only the best bidirectional best hit was retained per segment. Filtered results were used to calculate the average nucleotide identity (ANI), and alignment fraction (AF) as defined earlier. The distance (abbreviated Total Average Nucleotide Identity, or tANI) was calculated by using the formula: tANI = -LN(AF*ANI). The natural log of this value ensures that higher distance values correlate with genomes that have a lower ANI or AF; hence, being more dissimilar.

Bootstrap Replicates. After genomes were split into 1020 nucleotide segments, individual segments were chosen with replacement from this dataset, and used to create a new dataset. The new dataset was then compared against all other genomes as described above to create a bootstrapped distance matrix. These matrices are then used to infer their own trees. Those trees are then mapped onto the best tree.

Coverage and Percent Identity Cutoffs: The original percent identity and coverage cutoff values were chosen based on those laid down by (Varghese et al., 2015). Cutoff values were primarily tested within the *Aeromonas* clade. Average distance within the clade was measured over a range of cutoff values **(FigureS06)** and multiple potential cutoffs were tested against the jANI standard cutoffs of 70% identity and 70% length. By comparing the new potential cutoff values and the more standard 70 at 70 cutoff's ability to construct accurate phylogenetic trees and comparing to trees of the same dataset built

using more conventional methods, we concluded that 70 at 70 still produced the most accurate trees.

Phylogenies from Distances

Treebuilding from distance matrices was accomplished using the R packages Ape and Phangorn (Paradis et al., 2004; Schliep, 2011). The balanced minimum evolution algorithm as implemented in the FastME function of APE was used to generate phylogenies for each distance matrix (Desper and Gascuel, 2002). Parameters used were: nni = TRUE, spr= TRUE, tbr = TRUE. A "best tree" was calculated from the point estimate values (original DDH estimations in *is*DDH; the initial calculated distance matrix in tANI) and a collection of bootstrap topologies from the resampled matrices. Support values were mapped onto the best tree using the function *plotBS* in Phangorn (Schliep, 2011).

Checking Parameters

Bootstrap evaluation. Tree certainty scores were calculated using the IC/TC score calculation algorithm implemented in RAxML v8 (Salichos et al., 2014; Stamatakis, 2014). Tree distances were calculated using the R packages Ape (Paradis et al., 2004) and the treedist function of Phangorn (Schliep, 2011).

Residual Operating Characteristic Curve Analysis

A residual operating characteristic (ROC) curve was used to determine the optimal species cutoff for a single genome-to-genome distance calculation. Genomes from the

sets of *Aeromonas* and *Roseobacters* listed in the genome table were compiled, and matrices of both the distance and raw jANI were compiled from the set. The jANI values were used to delimit groups of species from the genomes selected, and was represented as a one for the true state. If a single calculation did not meet the cutoff value for a species (Richter and Rosselló-Móra, 2009), then the calculation had a zero for the true state.

True states and distance values were then compiled into a two-column data set. The R package pROC (Robin et al., 2011) allowed us to create a curve from the data, and then using methodology previously described (Youden, 1950) determined the best cutoff values for the given set of data such that true negatives and true positives based on the cutoff value were maximized.

Discussion

Success of treebuilding

The two novel methods we describe for generating supported phylogenies from whole-genome comparisons both demonstrated the capacity to match more sophisticated techniques.

The *is*DDH method revealed a more limited range to its utility than tANI. While the *is*DDH phylogeny did not match the canonical *Aeromonas* topology, the conflicts were noteworthy for being poorly supported or being located in a controversial clade that has seen many changes to its nomenclatures over time (Huys et al., 2001; Silver et al., 2011). While this does not excuse the performance of the method, it does explain such disagreements. At the family level in the Roseo dataset the isDDH method proved inadequate to matching the reference topology. Furthermore, at the order level, using the AeroOG set the method appears to have broken down completely, placing taxa in seeming random locations. We surmise that the isDDH treebuilding method we present here is of little value beyond the genus level. This agrees with the assertions of the GGDC's authors, who do not regard looking at deeper taxonomic ranks as productive use of their method **(**http://ggdc.dsmz.de/faq.php#qggdc17).

In contrast to the isDDH method, tANI showed negligible conflict with the reference at the genus level. Additionally, it agreed well at the family level in the Roseo set, identifying the same issues with *Ruegeria* and *Loktanella* paraphyly as the reference revealed. At the order level the relationships observed in the reference held true in the ANI-based tree. This phylogeny and the associated distance values also identified the *Succinovibrionaceae* as candidates for reclassification. While our testing is not

comprehensive we are of the opinion that this has demonstrated the suitability for using ANI to infer phylogenies to at least the family level and likely into higher ranks.

tANI is not affected by biases

The core of this work is predicated on the assumption that the genome as a whole conveys a significant amount of relevant information about the history of the organism. This assumption is broadly comparable to those made in using genomic content information to infer phylogeny and is subject to many of the same critiques (Wolf et al., 2002). There are two primary issues to consider.

First, in light of potentially rampant HGT, how much of a cell's genome will reflect a history of cell divisions rather than a composite of signals from the organism's recombination partners? Fortunately, this has a reasonable answer. Andam et al., has demonstrated that closely related taxa exchange genes more frequently than do more distantly related taxa (Andam and Gogarten, 2011). Importantly, the patterns of these exchanges mirror patterns of vertical inheritance. Thus, the signatures of HGT that might be picked up in this method would also largely mirror the vertical descent of the taxon. How much this applies to deeper taxonomic ranks, however, is unfortunately not so certain. It is possible that the flows of gene-sharing that unite and divide such close relatives as *Escherichia* and *Salmonella* may not behave in the same way with more distant relationships.

Second, two organisms failing to share highly similar genome content is hardly proof of their being highly diverged. Two instructive examples are the obligate intracellular parasites and the Halobacteria. The highly reduced and streamlined genomes of the *Wolbachia*, *Ricketsialles*, and others initially led phylogenetic methods to conclude

they formed a single clade because of their long branches and tiny genomes (Canbäck et al., 2004). It was not until more sophisticated approaches developed that it was realized there have been multiple convergent emergences of intracellularity (Herbeck et al., 2005). Likewise, the Halobacteria, colloquially known as the Haloarchaea as they are not Bacteria, were placed at the base of the entire Archaeal domain using gene content analyses (Wolf et al., 2002) when their currently accepted position is far more derived and sister to the Methanomicrobia (Delsuc et al., 2005). While this concern does not have an easy answer we were pleased to discover that our ANI-extension is at least somewhat robust to these issues. As our experiments with the Frankia set demonstrate, a nearly 3-fold range of genome size nor a 15% GC-content range had any discernable impact on our derived phylogeny. While far from proving such issues cannot derail the method it is affirming to see the boundaries pushed back so far.

If these limits prove to extend no deeper one could restrict their analysis to so-called "free-living" organisms. This would avoid the issue of streamlined obligate intracellular taxa and the disparity in genome size that entails. However, the Halobacteria's genomic signature is the result of its existence in, as well as strategy for dealing with, high-salt environments, rather than intracellularity. Restricting analyses to only "free-living" taxa creates a number of philosophical dilemmas, not the least of which is what one defines as free living, that we do not have any easy solutions for. We would think it dubious to completely exclude organisms from study by this and related methods solely because of their lifestyles. However, we would caution against blindly trusting the methodology to give a reasonable answer without considering the peculiarities of the data. Which, is perhaps the exact behavior we are seeking to enable by extending ANI in the fashion.

Misclassified taxa

Results from our methods show that there is a clear separation of the *Loktanella* and *Ruegeria* genera into multiple separate clades, however *Loktanella* is significantly more fragmented (**Figure05**). The conclusion that these classifications should be re-described is supported by results from previous literature on *Ruegeria.* Clustering of the genus in previous studies would always group *Ruegeria* strains together; however these studies often only included one or two members of the genus, or had poorly selected outgroups and a lack of resolution outside the genus. (Martens et al., 2006; Yi et al., 2007). Studies that did include more members from the genus *Ruegeria* have often not reported support values for the clades they are present in (Breider et al., 2014; Park and Yoon, 2012; Vandecandelaere et al., 2008). *Loktanella* may also require a revisit, as previous literature would suggest that the results of the phylogeny in (**Figure05**) are more reflective of the actual phylogeny**.** Previous classification studies have often used poor support values, failed to report support values, or use a set of genomes without proper resolution to claim that a number of strains belong to the genus (Lee, 2012; Moon et al., 2010; Tsubouchi et al., 2013; Van Trappen et al., 2004). These studies, often reporting <60% support values for nodes grouping the *Loktanella* as a single clade (Lee, 2012; Moon et al., 2010; Tsubouchi et al., 2013; Van Trappen et al., 2004), suggest that our results may be indicative of the true nature of the *Loktanella* group.

Deeper taxonomic ranks

A pleasing side effect to creating a single unified distance measure from our ANI-extension and testing our treebuilding on deeper taxonomic ranks was the opportunity to investigate deeper taxonomic cutoffs. In the same sense that ANI and GGDC have been

used to delimit species, and in the case of GGDC strains, we examined whether our distance value could discriminate genus, family, and orders. While our test sets are not exhaustive the results were promising. Genus assignments were achieved at a rate of ~10% false positives and false negatives at ~1%. At family the false positives remained roughly unchanged but the false negatives declined to ~14%. While not as sterling as the species discrimination (< 1% for both), their specificities match and sensitivities nearly double enzymatic rapid diagnostic testing for influenza (https://www.cdc.gov/flu/professionals/diagnosis/rapidlab.htm). We are optimistic that further examination using a more comprehensive taxon set will support the general utility of this distance as a generic taxonomic delimiter. As with previous iterations of ANI different groups may require specific considerations outside of a one cutoff fits all mold.

Overall, we feel we have identified several valuable extensions to whole-genome comparison data that is being routinely generated by researchers as a matter of course. The ability to produce viable and statistically supported phylogenies offers the possibility for researchers to save time on more complex phylogenetics. Simultaneously, it offers the hope of a more sophisticated and reliable result than simply creating a 16S rDNA parsimony tree from assembled and likely chimeric ORFs. Furthermore, the possibility that the ANI method can differentiate deeper taxonomic relationships offers the glimpse of hope that it may be able to bring the same light to the delimitation of high taxonomic ranks as ANI and GGDC has brought to species and strain classification. Finally, perhaps the greatest benefit of these developments, and worth reiterating, is that for many researchers the input data, or a close cousin, is already being generated *de rigeur* in much

the same way that sequencing a 16S fragment and throwing together a quick parsimony tree became a requirement to publishing on a new isolate years ago. We offer new option to make use of that already present data in constructive ways.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi:10.1093/nar/25.17.3389.

Andam, C. P., and Gogarten, J. P. (2011). Biased gene transfer and its implications for the concept of lineage. *Biol. Direct* 6, 47. doi:10.1186/1745-6150-6-47.

Auch, A. F., Henz, S. R., Holland, B. R., and Göker, M. (2006). Genome BLAST distance phylogenies inferred from whole plastid and whole mitochondrion genome sequences. *BMC Bioinformatics* 7, 350. doi:10.1186/1471-2105-7-350.

Auch, A. F., Jan, M. von, Klenk, H.-P., and Göker, M. (2010a). Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand. Genomic Sci.* 2, 117–134. doi:10.4056/sigs.531120.

Auch, A. F., Klenk, H.-P., and Göker, M. (2010b). Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. *Stand. Genomic Sci.* 2, 142–148. doi:10.4056/sigs.541628.

Breider, S., Scheuner, C., Schumann, P., Fiebig, A., Petersen, J., Pradella, S., et al. (2014). Genome-scale data suggest reclassifications in the Leisingera-Phaeobacter cluster including proposals for Sedimentitalea gen. nov. and Pseudophaeobacter gen. nov. *Front. Microbiol.* 5. doi:10.3389/fmicb.2014.00416.

Canbäck, B., Tamas, I., and Andersson, S. G. E. (2004). A Phylogenomic Study of Endosymbiotic Bacteria. *Mol. Biol. Evol.* 21, 1110–1122. doi:10.1093/molbev/msh122.

Chernomor, O., von Haeseler, A., and Minh, B. Q. (2016). Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Syst. Biol.* 65, 997–1008. doi:10.1093/sysbio/syw037.

Colston, S. M., Fullmer, M. S., Beka, L., Lamy, B., Gogarten, J. P., and Graf, J. (2014). Bioinformatic Genome Comparisons for Taxonomic and Phylogenetic Assignments Using Aeromonas as a Test Case. *mBio* 5, e02136-14. doi:10.1128/mBio.02136-14.

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* 6, 361. doi:10.1038/nrg1603.

Desper, R., and Gascuel, O. (2002). Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum-Evolution Principle. *J. Comput. Biol.* 9, 687–705. doi:10.1089/106652702761034136.

Fullmer, M. S., Soucy, S. M., Swithers, K. S., Makkay, A. M., Wheeler, R., Ventosa, A., et al. (2014). Population and genomic analysis of the genus Halorubrum. *Extreme Microbiol.* 5, 140. doi:10.3389/fmicb.2014.00140.

Gibbon, F., T, S., and House, C. H. (1999). Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.* 27, 4218–4222. doi:10.1093/nar/27.21.4218.

Gogarten, J. P., Doolittle, W. F., and Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* 19, 2226–2238.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., and Tiedje, J. M. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* 57, 81–91. doi:10.1099/ijs.0.64483-0.

Haeseler, A. von, Minh, B. Q., Jermiin, L. S., Kalyaanamoorthy, S., and Wong, T. K. F. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587. doi:10.1038/nmeth.4285.

Hahnke, R. L., Meier-Kolthoff, J. P., García-López, M., Mukherjee, S., Huntemann, M., Ivanova, N. N., et al. (2016). Genome-Based Taxonomic Classification of Bacteroidetes. *Front. Microbiol.* 7. doi:10.3389/fmicb.2016.02003.

Henz, S. R., Huson, D. H., Auch, A. F., Nieselt-Struwe, K., and Schuster, S. C. (2005). Whole-genome prokaryotic phylogeny. *Bioinformatics* 21, 2329–2335. doi:10.1093/bioinformatics/bth324.

Herbeck, J. T., Degnan, P. H., and Wernegreen, J. J. (2005). Nonhomogeneous Model of Sequence Evolution Indicates Independent Origins of Primary Endosymbionts Within the Enterobacteriales (γ-Proteobacteria). *Mol. Biol. Evol.* 22, 520–532. doi:10.1093/molbev/msi036.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Le, S. V. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* doi:10.1093/molbev/msx281.

Huys, G., Kämpfer, P., and Swings, J. (2001). New DNA-DNA hybridization and phenotypic data on the species Aeromonas ichthiosmia and Aeromonas allosaccharophila:a. Ichthiosmia schubert et al. 1990 is a later synonym of a. veronii hickman-brenner et al. 1987. *Syst. Appl. Microbiol.* 24, 177–182. doi:10.1078/0723-2020-00038.

Jain, R., Rivera, M. C., and Lake, J. A. (1999). Horizontal gene transfer among genomes: The complexity hypothesis. *Proc. Natl. Acad. Sci.* 96, 3801–3806. doi:10.1073/pnas.96.7.3801.

Konstantinidis, K. T., Ramette, A., and Tiedje, J. M. (2006). The bacterial species definition in the genomic era. *Philos. Trans. R. Soc. B Biol. Sci.* 361, 1929–1940. doi:10.1098/rstb.2006.1920.

Konstantinidis, K. T., and Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 102, 2567–2572. doi:10.1073/pnas.0409727102.

Krajewski, C., and Dickerman, A. W. (1990). Bootstrap Analysis of Phylogenetic Trees Derived from DNA Hybridization Distances. *Syst. Biol.* 39, 383–390. doi:10.2307/2992358.

Lee, S. D. (2012). Loktanella tamlensis sp. nov., isolated from seawater. *Int. J. Syst. Evol. Microbiol.* 62, 586–590. doi:10.1099/ijs.0.029462-0.

Martens, T., Heidorn, T., Pukall, R., Simon, M., Tindall, B. J., and Brinkhoff, T. (2006). Reclassification of Roseobacter gallaeciensis Ruiz-Ponte et al. 1998 as Phaeobacter gallaeciensis gen. nov., comb. nov., description of Phaeobacter inhibens sp. nov., reclassification of Ruegeria algicola (Lafay et al. 1995) Uchino et al. 1999 as Marinovum algicola gen. nov., comb. nov., and emended descriptions of the genera Roseobacter, Ruegeria and Leisingera. *Int. J. Syst. Evol. Microbiol.* 56, 1293–1304. doi:10.1099/ijs.0.63724-0.

Martin-Carnahan, A., and Joseph, S. W. (2015). "Aeromonadales ord. nov," in *Bergey's Manual of Systematics of Archaea and Bacteria* (John Wiley & Sons, Ltd). doi:10.1002/9781118960608.obm00093.

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14, 60. doi:10.1186/1471-2105-14-60.

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P., and Göker, M. (2014). Highly parallelized inference of large genome-based phylogenies. *Concurr. Comput. Pract. Exp.* 26, 1715–1729. doi:10.1002/cpe.3112.

Moon, Y. G., Seo, S. H., Lee, S. D., and Heo, M. S. (2010). Loktanella pyoseonensis sp. nov., isolated from beach sand, and emended description of the genus Loktanella. *Int. J. Syst. Evol. Microbiol.* 60, 785–789. doi:10.1099/ijs.0.011072-0.

Mukherjee, S., Seshadri, R., Varghese, N. J., Eloe-Fadrosh, E. A., Meier-Kolthoff, J. P., Göker, M., et al. (2017). 1,003 reference genomes of bacterial and archaeal

isolates expand coverage of the tree of life. *Nat. Biotechnol.* 35, 676–683. doi:10.1038/nbt.3886.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi:10.1093/molbev/msu300.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 17, 132. doi:10.1186/s13059-016-0997-x.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20, 289–290. doi:10.1093/bioinformatics/btg412.

Park, S., and Yoon, J.-H. (2012). Ruegeria arenilitoris sp. nov., isolated from the seashore sand around a seaweed farm. *Antonie Van Leeuwenhoek* 102, 581–589. doi:10.1007/s10482-012-9753-8.

Peeters, C., Meier-Kolthoff, J. P., Verheyde, B., De Brandt, E., Cooper, V. S., and Vandamme, P. (2016). Phylogenomic Study of Burkholderia glathei-like Organisms, Proposal of 13 Novel Burkholderia Species and Emended Descriptions of Burkholderia sordidicola, Burkholderia zhejiangensis, and Burkholderia grimmiae. *Front. Microbiol.* 7. doi:10.3389/fmicb.2016.00877.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing Available at: http://www.R-project.org.

Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci.* 106, 19126–19131. doi:10.1073/pnas.0906412106.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12, 77. doi:10.1186/1471-2105-12-77.

Salichos, L., Stamatakis, A., and Rokas, A. (2014). Novel Information Theory-Based Measures for Quantifying Incongruence among Phylogenetic Trees. *Mol. Biol. Evol.* 31, 1261–1271. doi:10.1093/molbev/msu061.

Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics* 27, 592–593. doi:10.1093/bioinformatics/btq706.

Silver, A. C., Williams, D., Faucher, J., Horneman, A. J., Gogarten, J. P., and Graf, J. (2011). Complex Evolutionary History of the Aeromonas veronii Group

Revealed by Host Interaction and DNA Sequence Data. *PLoS ONE* 6, e16751. doi:10.1371/journal.pone.0016751.

Stackebrandt, E., and Goebel, B. M. (1994). Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int. J. Syst. Bacteriol.* 44, 846–849. doi:10.1099/00207713-44-4-846.

Stackebrandt, E., and Hespell, R. B. (2006). "The Family Succinivibrionaceae," in *The Prokaryotes*, eds. M. D. P. Dr, S. Falkow, E. Rosenberg, K.-H. Schleifer, and E. Stackebrandt (Springer New York), 419–429. doi:10.1007/0-387-30743-5_20.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi:10.1093/bioinformatics/btu033.

Tsubouchi, T., Shimane, Y., Mori, K., Miyazaki, M., Tame, A., Uematsu, K., et al. (2013). Loktanella cinnabarina sp. nov., isolated from a deep subseafloor sediment, and emended description of the genus Loktanella. *Int. J. Syst. Evol. Microbiol.* 63, 1390–1395. doi:10.1099/ijs.0.043174-0.

Van Trappen, S., Mergaert, J., and Swings, J. (2004). Loktanella salsilacus gen. nov., sp. nov., Loktanella fryxellensis sp. nov. and Loktanella vestfoldensis sp. nov., new members of the Rhodobacter group, isolated from microbial mats in Antarctic lakes. *Int. J. Syst. Evol. Microbiol.* 54, 1263–1269. doi:10.1099/ijs.0.03006-0.

Vandecandelaere, I., Nercessian, O., Segaert, E., Achouak, W., Faimali, M., and Vandamme, P. (2008). Ruegeria scottomollicae sp. nov., isolated from a marine electroactive biofilm. *Int. J. Syst. Evol. Microbiol.* 58, 2726–2733. doi:10.1099/ijs.0.65843-0.

Varghese, N. J., Mukherjee, S., Ivanova, N., Konstantinidis, K. T., Mavrommatis, K., Kyrpides, N. C., et al. (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Res.* 43, 6761–6771. doi:10.1093/nar/gkv657.

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., and Koonin, E. V. (2002). Genome trees and the tree of life. *Trends Genet.* 18, 472–479. doi:10.1016/S0168-9525(02)02744-0.

Yi, H., Lim, Y. W., and Chun, J. (2007). Taxonomic evaluation of the genera Ruegeria and Silicibacter: a proposal to transfer the genus Silicibacter Petursdottir and Kristjansson 1999 to the genus Ruegeria Uchino et al. 1999. *Int. J. Syst. Evol. Microbiol.* 57, 815–819. doi:10.1099/ijs.0.64568-0.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer* 3, 32–35. doi:10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

Chapter 3.1.1 Figures & Tables

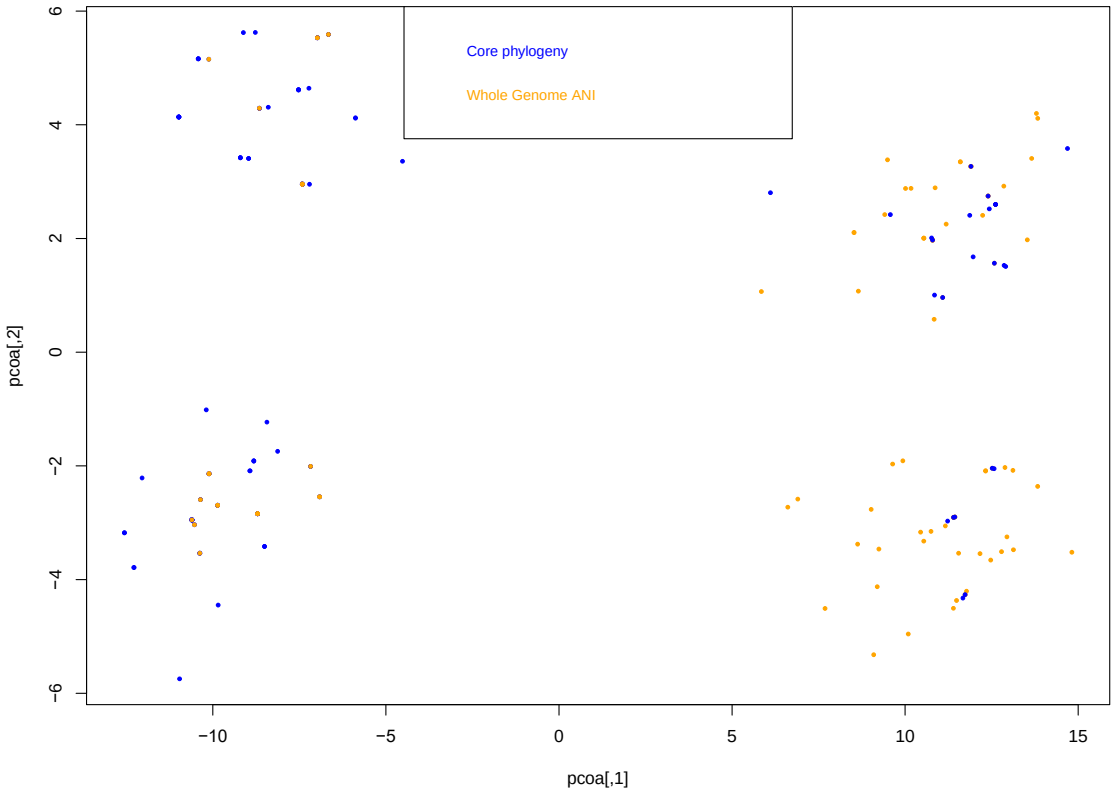


Figure01. PCoA plot of the bootstrapped tree distances of the mBio and ANI-methods. Distances were calculated using Robinson-Foulds distances in all-vs-all fashion. The support sets overlap in every cluster, strongly suggesting that the two methods are capturing the same topologies.

Figure02. ANI-method phylogeny of the Frankia dataset rooted in accordance with NCBI's taxonomy data. Tip labels are colored by genome size. These results illustrate that large difference the size of genomes does not appear to bias the results of the tANI-method.

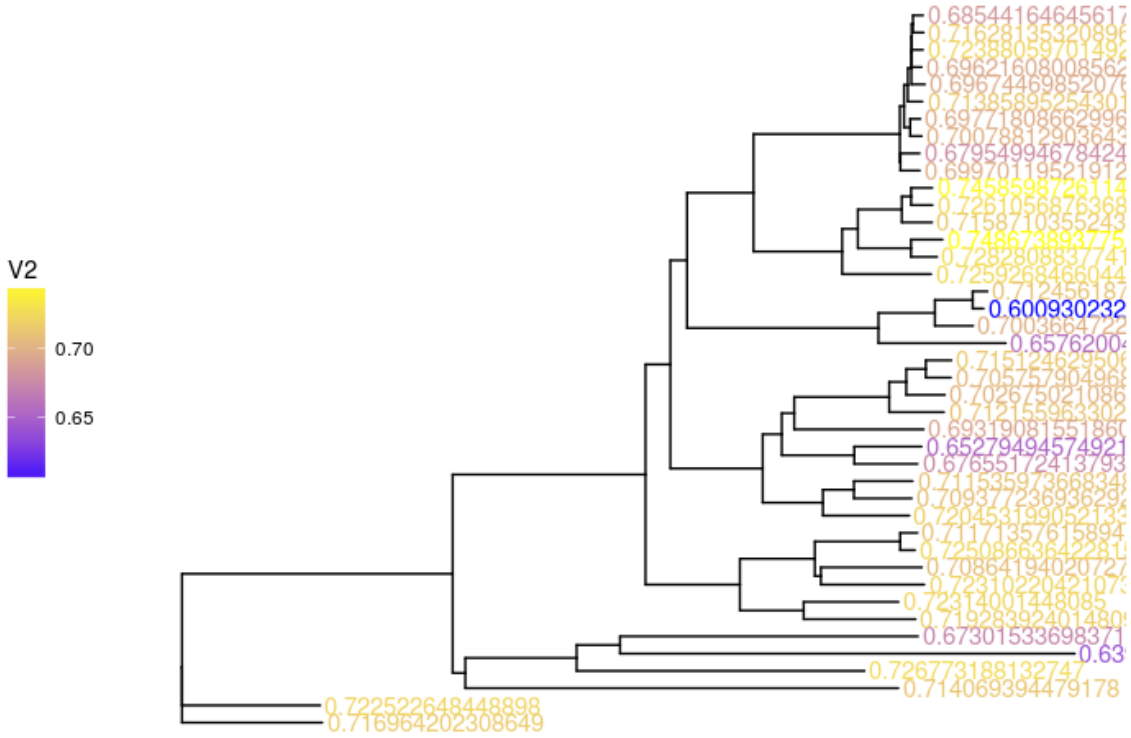


Figure03. ANI-method phylogeny of the Frankia dataset rooted in accordance with NCBI's taxonomy data. Genomic %GC are both substituted for the tip labels and also provide the color-coding. These results illustrate that large difference in %GC does not appear to bias the results of the tANI-method.

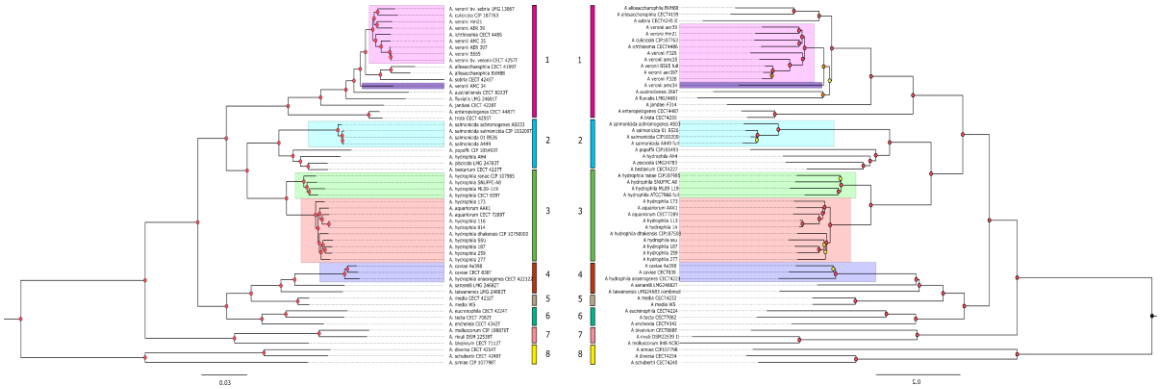


Figure04. Comparison of mBio Extended Core Phylogeny, inferred using Approximate Maximum-likelihood (Colston et al., 2014), and tANI, inferred using Fast Minimum Evolution (Desper and Gascuel, 2002). The two topologies are nearly identical with very similar support for their shared nodes.

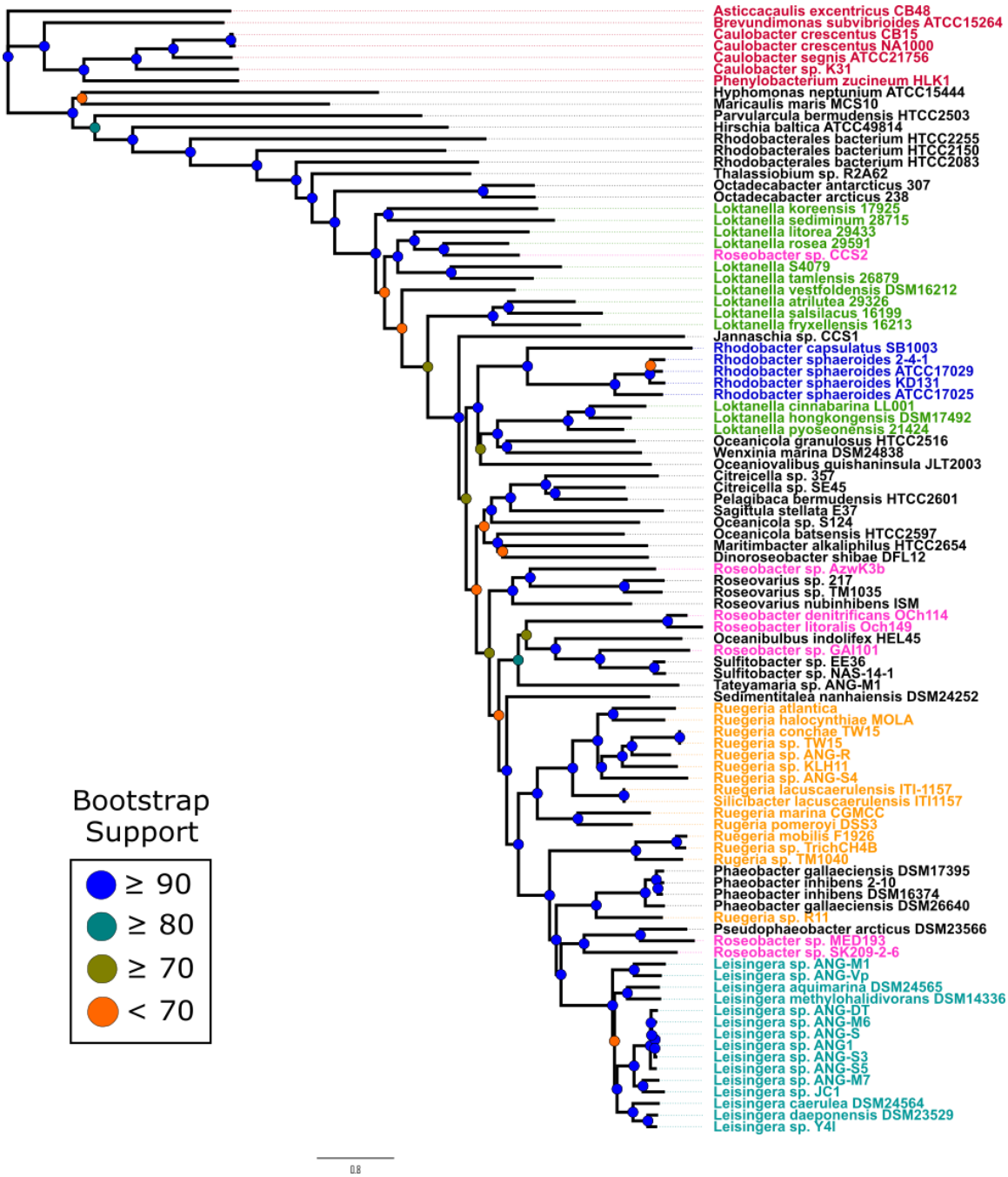


Figure05a. Phylogeny of the Roseobacteriales dataset using the tANI treebuilding method. Genera are color-coded to highlight misidentified taxa.

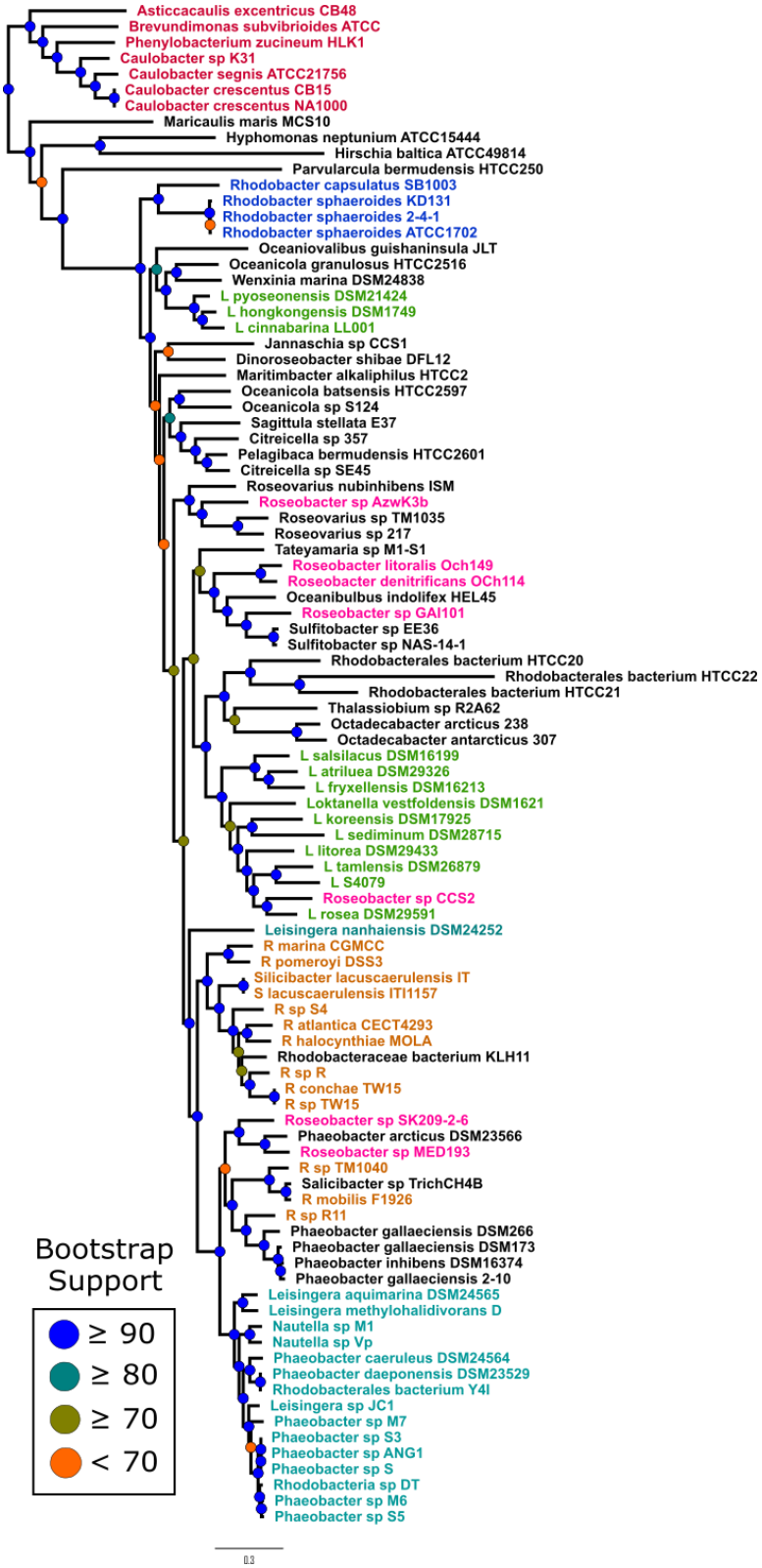


Figure05b. Phylogeny of the Roseobacteriales dataset using the multi-gene phylogeny from Collins et al., (2015). Genera are color-coded to highlight misidentified taxa.

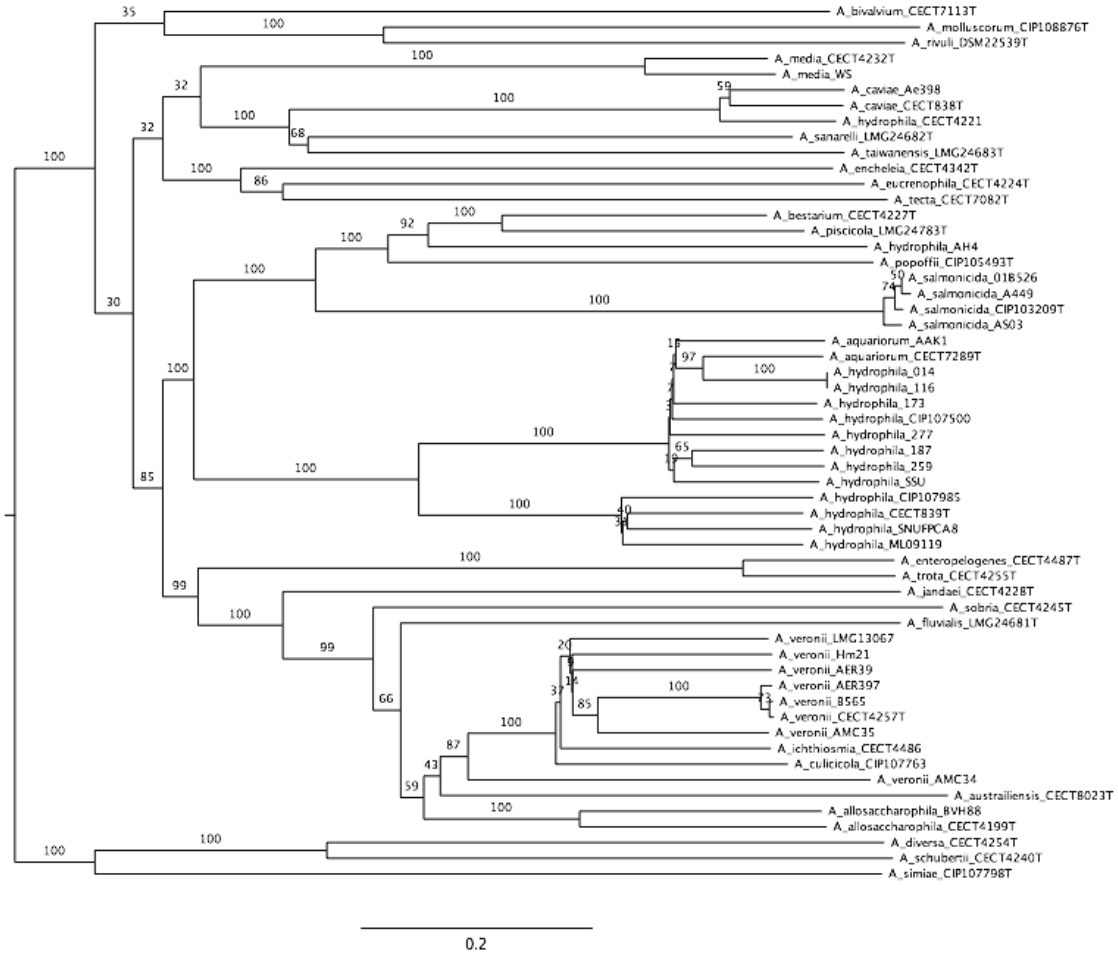


Figure06. Phylogeny of the mBio dataset inferred using the *is*DDH method. The tree's topology is largely in agreement with the reference. Support values are mostly strong, although several deeper nodes are low.

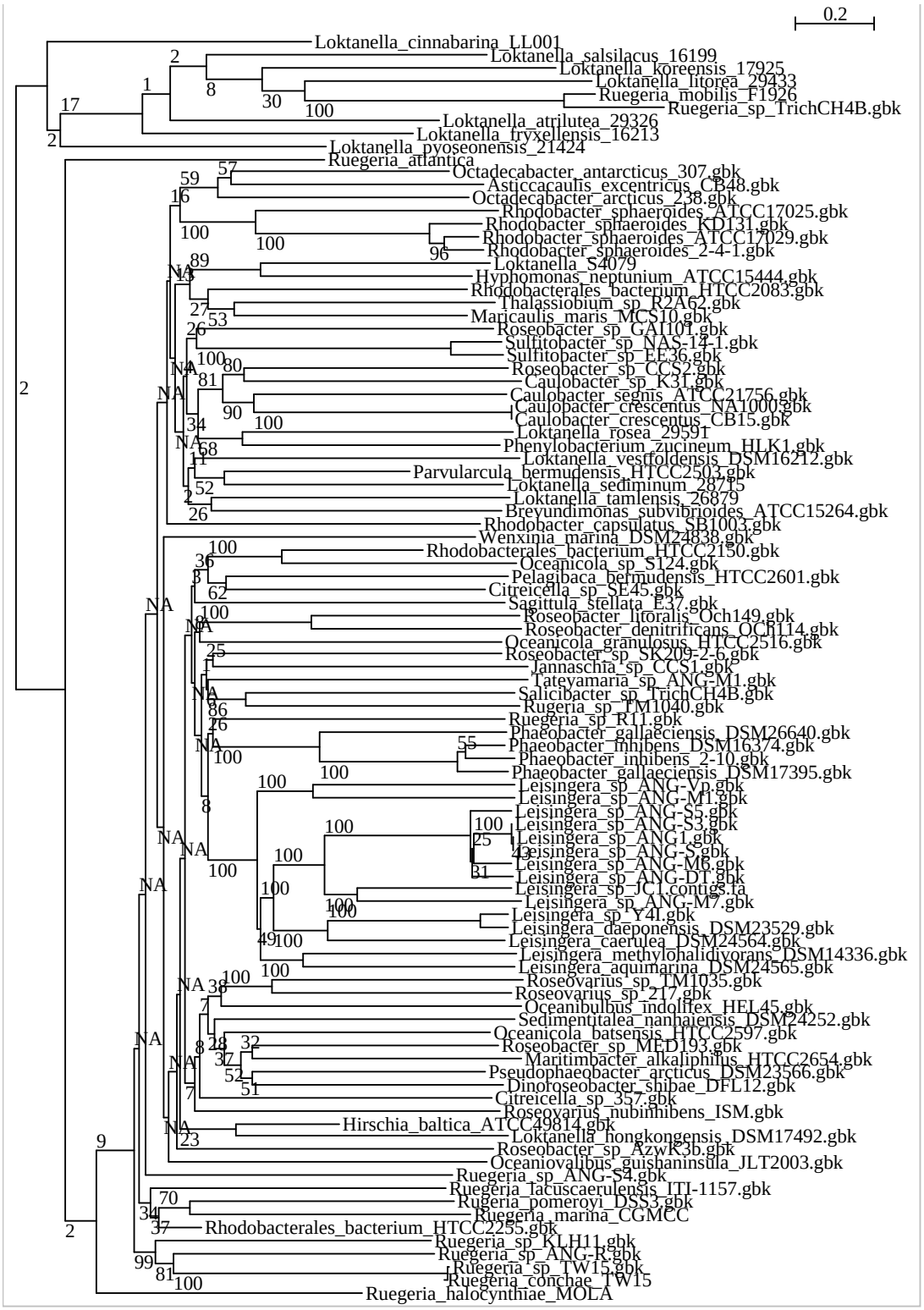


Figure07. Phylogeny of the Roseo dataset inferred using the *is*DDH method. The tree's topology very poorly matches the MLSA reference. Its bootstrap support also fails to significantly support many of the nodes.

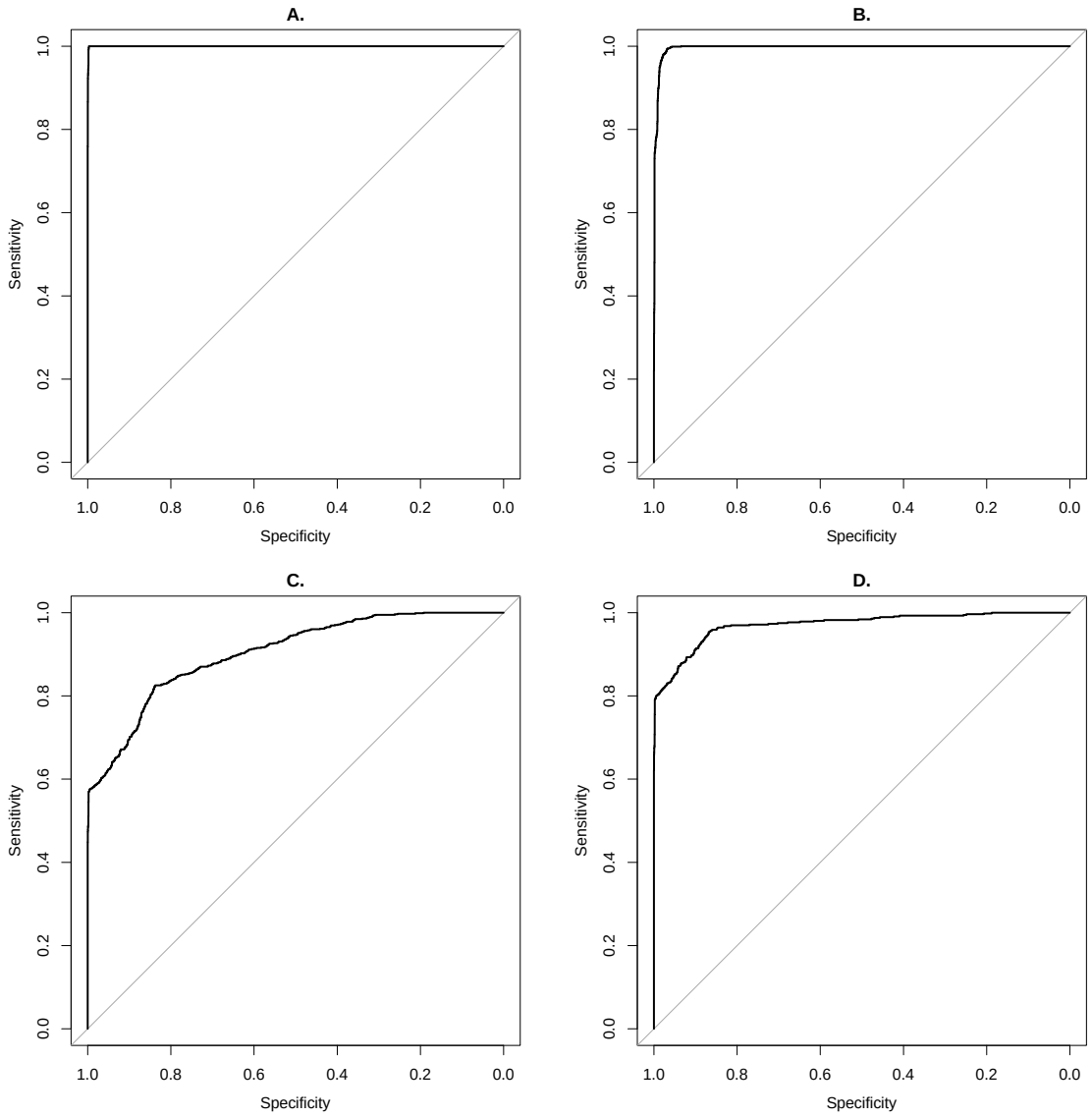
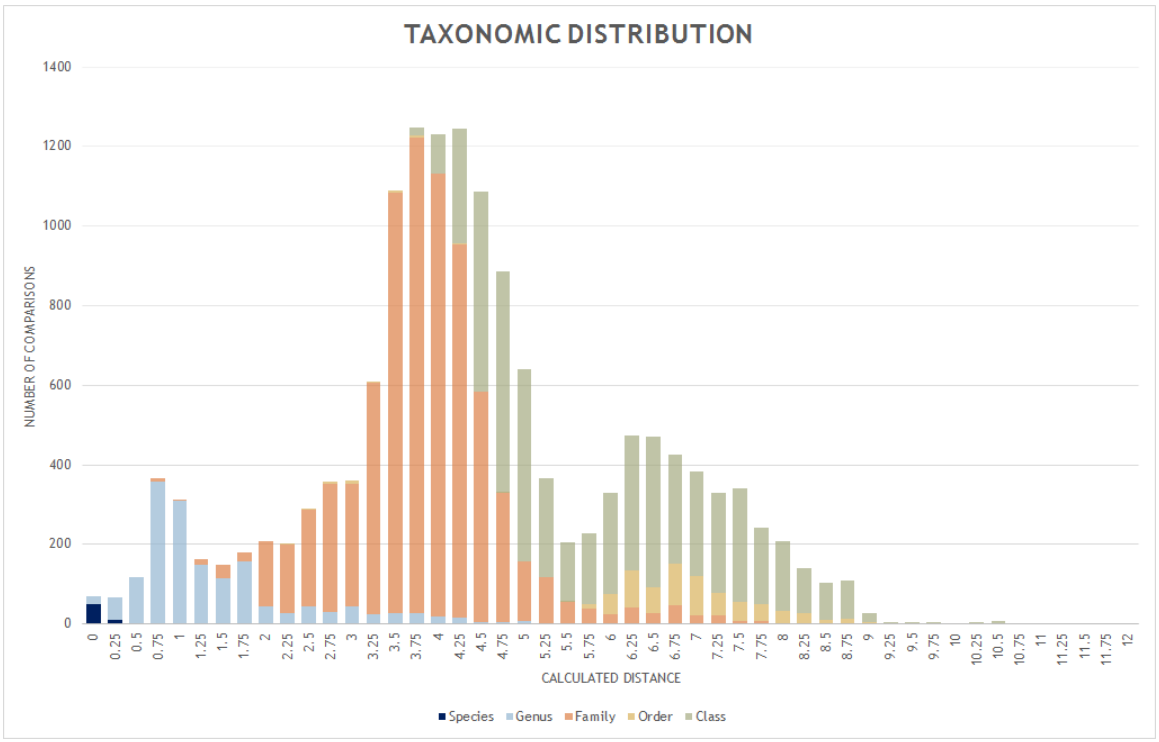


Figure08. Response Operator Curves reporting the sensitivity and specificity of the ANI-distance at discriminating species relationships. Panel A shows the union of the mBio and Roseo datasets against accepted nomenclature (specificity of 99.98%, and sensitivity of 99.20%). Panels B & C shows the mBio (specificity of 96.68%, and sensitivity of 97.97%) and Roseo (specificity of 83.78%, and sensitivity of 80.09%) datasets respectively, demonstrating that the two genera are not equally easy to discriminate. Panel D shows the Roseo set after reclassifying taxa (specificity of 83.31%, and sensitivity of 99.13%). It is noteworthy that the specificity spikes from 80% to 99%, further improving the performance of our method.

A.



B.

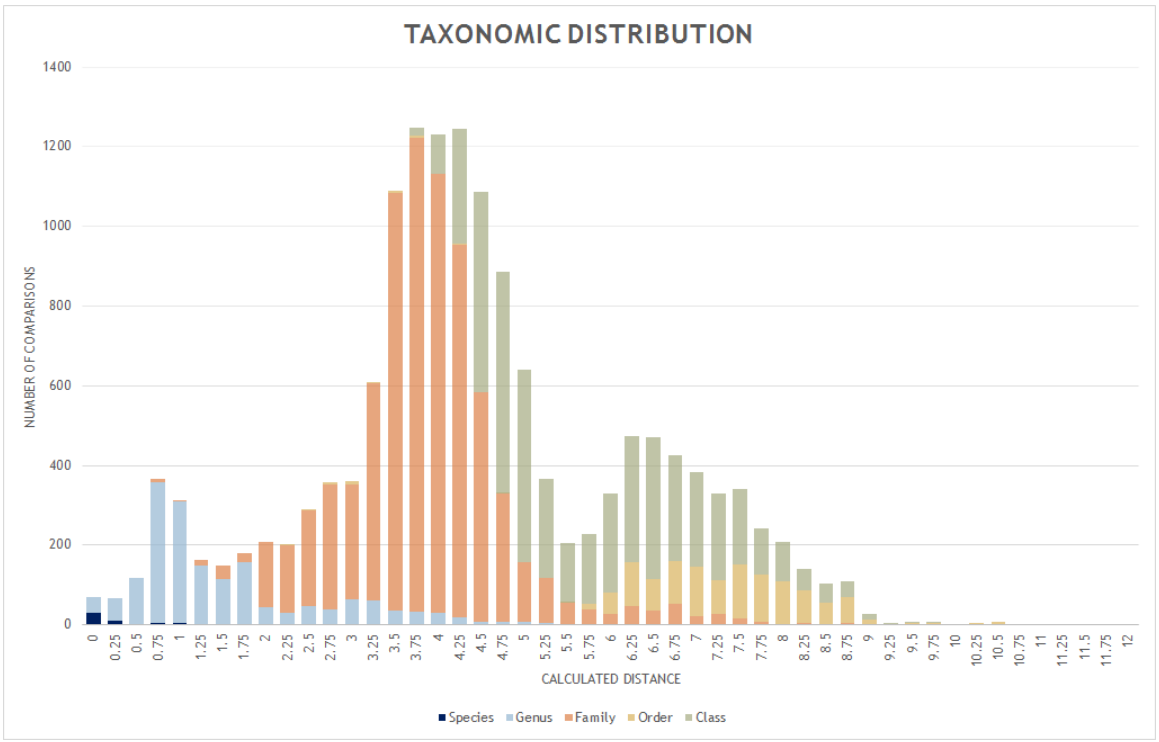


Figure09. Histograms relating the numbers of taxonomic rank comparisons in our datasets as functions of our ANI-distance values. Panel A displays uncorrected taxonomy as derived from NCBI. Panel B displays the distribution after identification of

misclassified taxa and re-categorization along the lines suggested in the reference phylogenies.

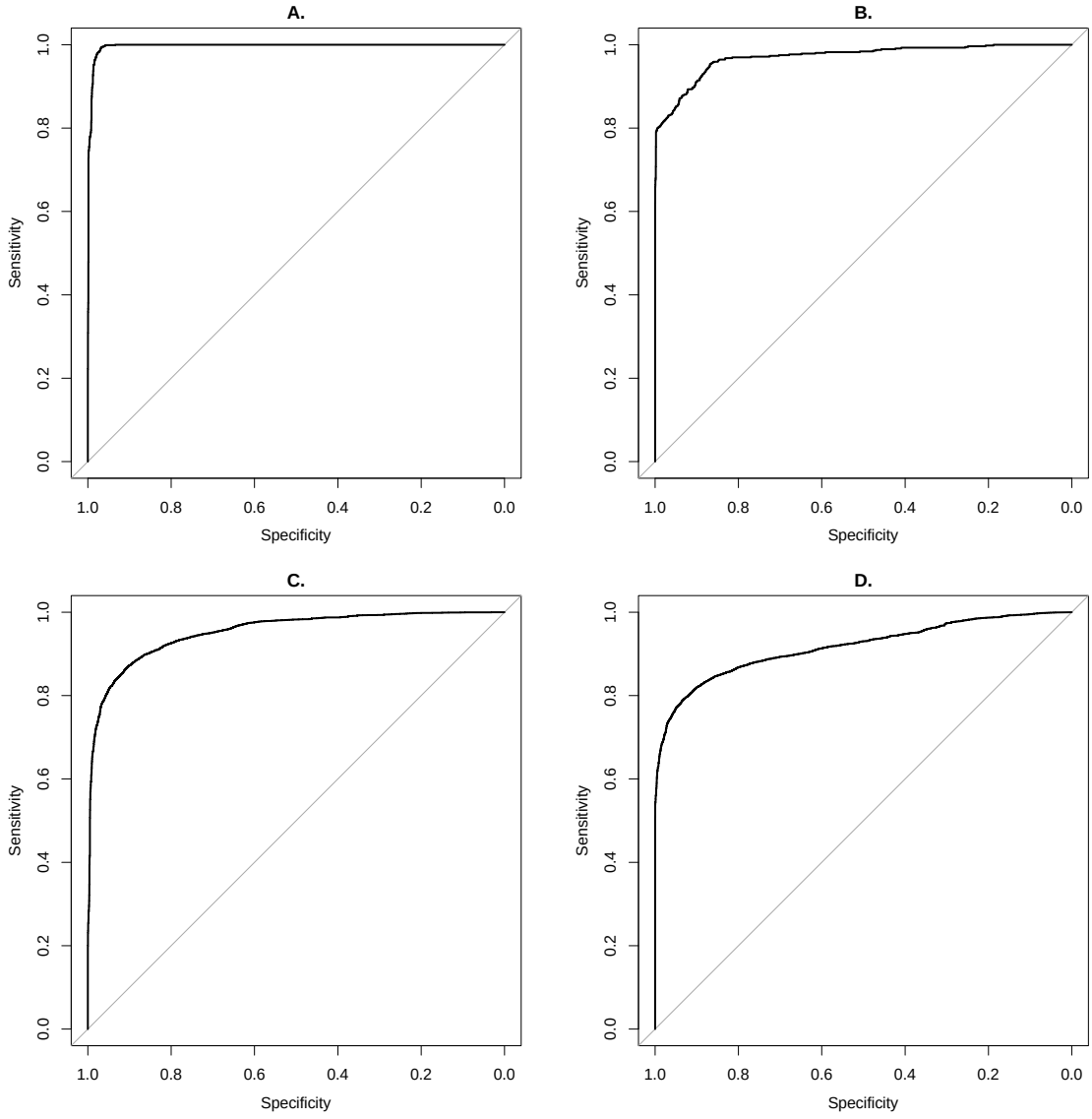


Figure10. Response Operator Curves reporting the sensitivity and specificity of the ANI-distance at discriminating deeper taxonomic relationships. Panel A shows the AeroOG dataset at the genus level. Panel B is the Roseo dataset, also at the genus level. Specificities (96.7% and 83.3%) and sensitivities (98.0% and 99.1%) are varied but generally high. Panel C shows our combined datasets at the family level. The family relationships maintain an ability to discriminate between classifications at rate close to the genus data (90.7% specificity and sensitivity of 86.5%). Panel D displays the combined data at order level. While order level specificity was high (94.2%) its sensitivity was only 71.4%, suggesting the method is breaking down and losing the ability to discriminate.

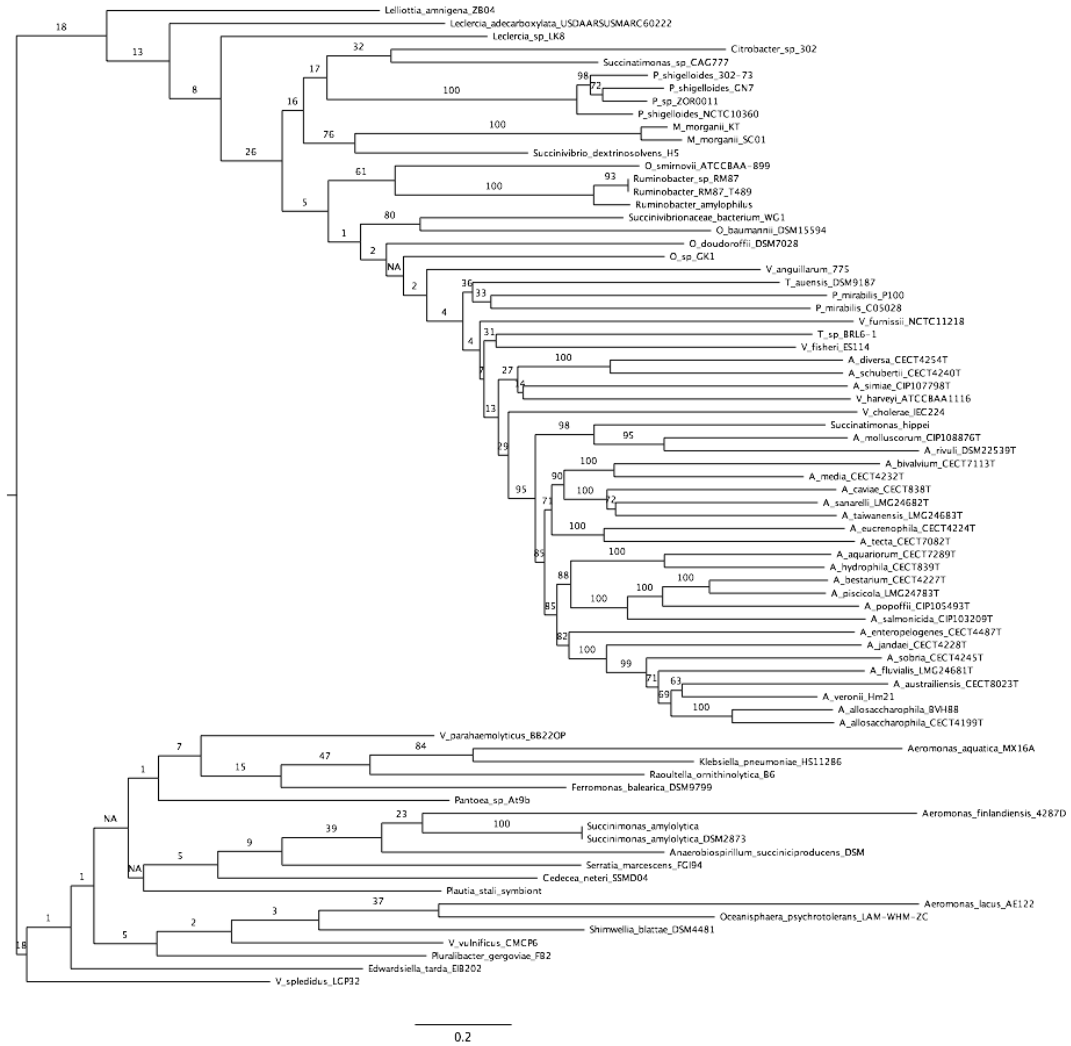
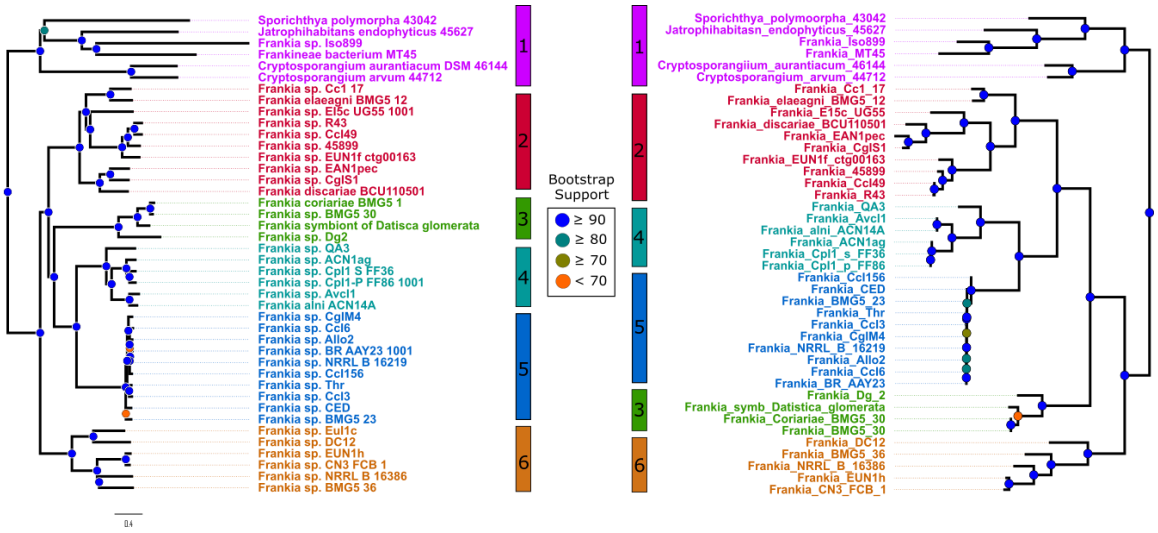


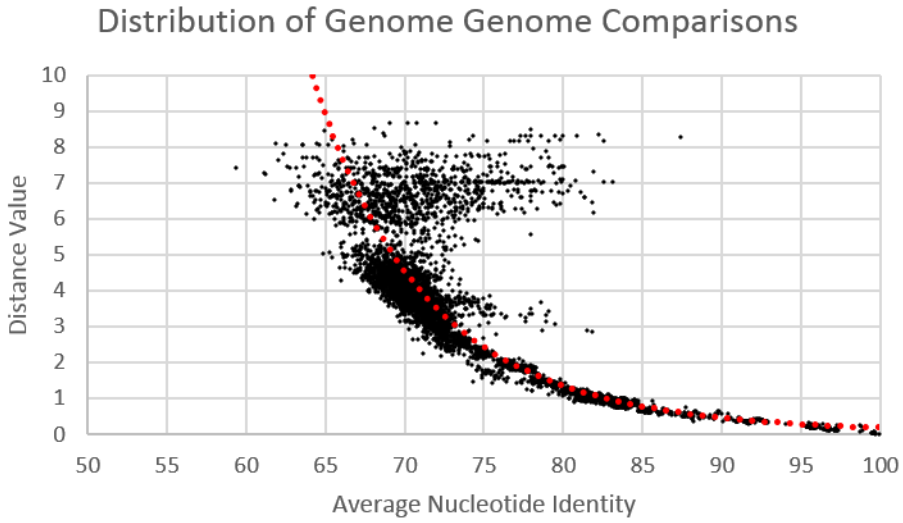
Figure11. Phylogeny of the AeroOG dataset inferred using the *is*DDH method. The tree's topology is hard to discern from random at a glance. The bootstrap supports are also broadly very poor. Combined with the weaknesses exhibited in Figures4&5, this suggests the *is*DDH method does not work.

Table01. Average IC values derived from mapping bootstrap sets onto best trees.

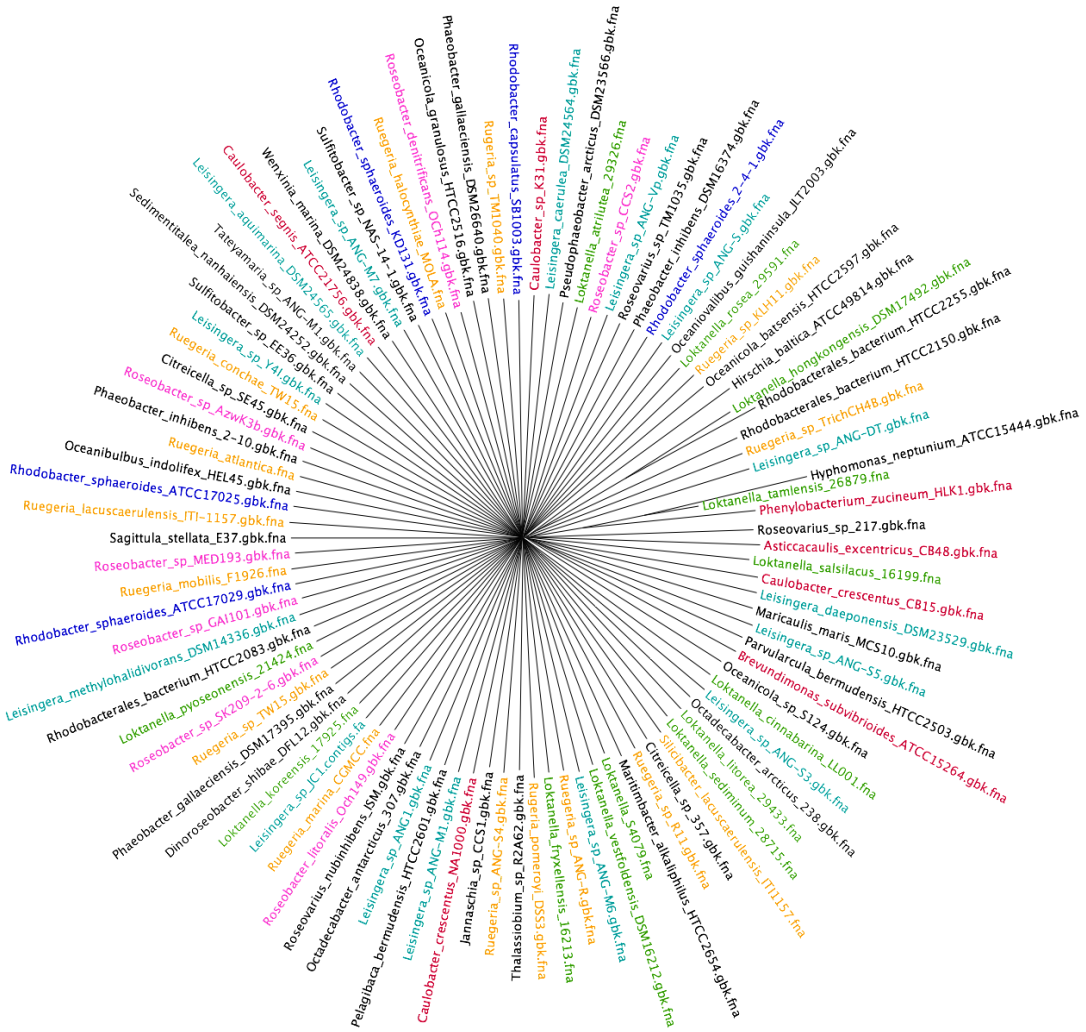
Dataset	Tree	Bootstrap Set	Avg. IC
Aero56	ANI	ANI	0.861751
	ANI	MLSA	0.32689
	ANI	core	0.608803
	MLSA	MLSA	0.652065
	core	ANI	0.610512
	core	core	0.874098
	GBDP_565	MLSA	0.014001
	GBDP_565	core	-0.014667
	mashtree	ANI	0.472861
	mashtree	MLSA	0.282831
Roseos	ANI	ANI	0.830630
	ANI	MLSA	0.386409
	MLSA	MLSA	0.803819
	mashtree	MLSA	0.250201
	mashtree	ANI	0.229349
	GBDP_569	ANI	-0.187205
	GBDP_569	MLSA	-0.225330



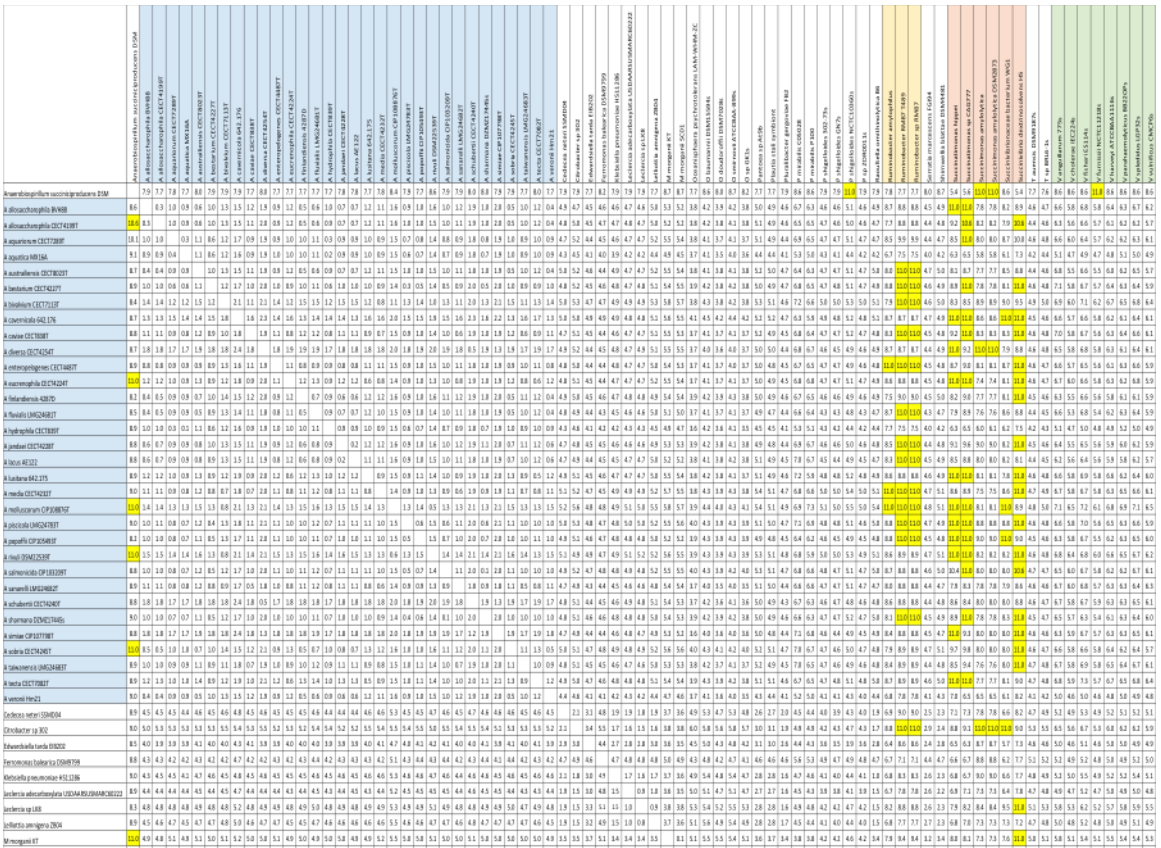
FigureS01. *Frankia* ANI-method phylogeny compared against *Frankia* MLSA phylogeny. Left tree is ANI-method and right tree is MLSA codon model.



FigureS02. tANI distance value as a function of uncorrected jSpecies ANI value. This "tornado" configuration illustrates how jSpecies ANI begins to enter saturation by approximately 87%. This saturation is a function of declining AF values and sequence saturation. The red dashed line portrays a power trendline.

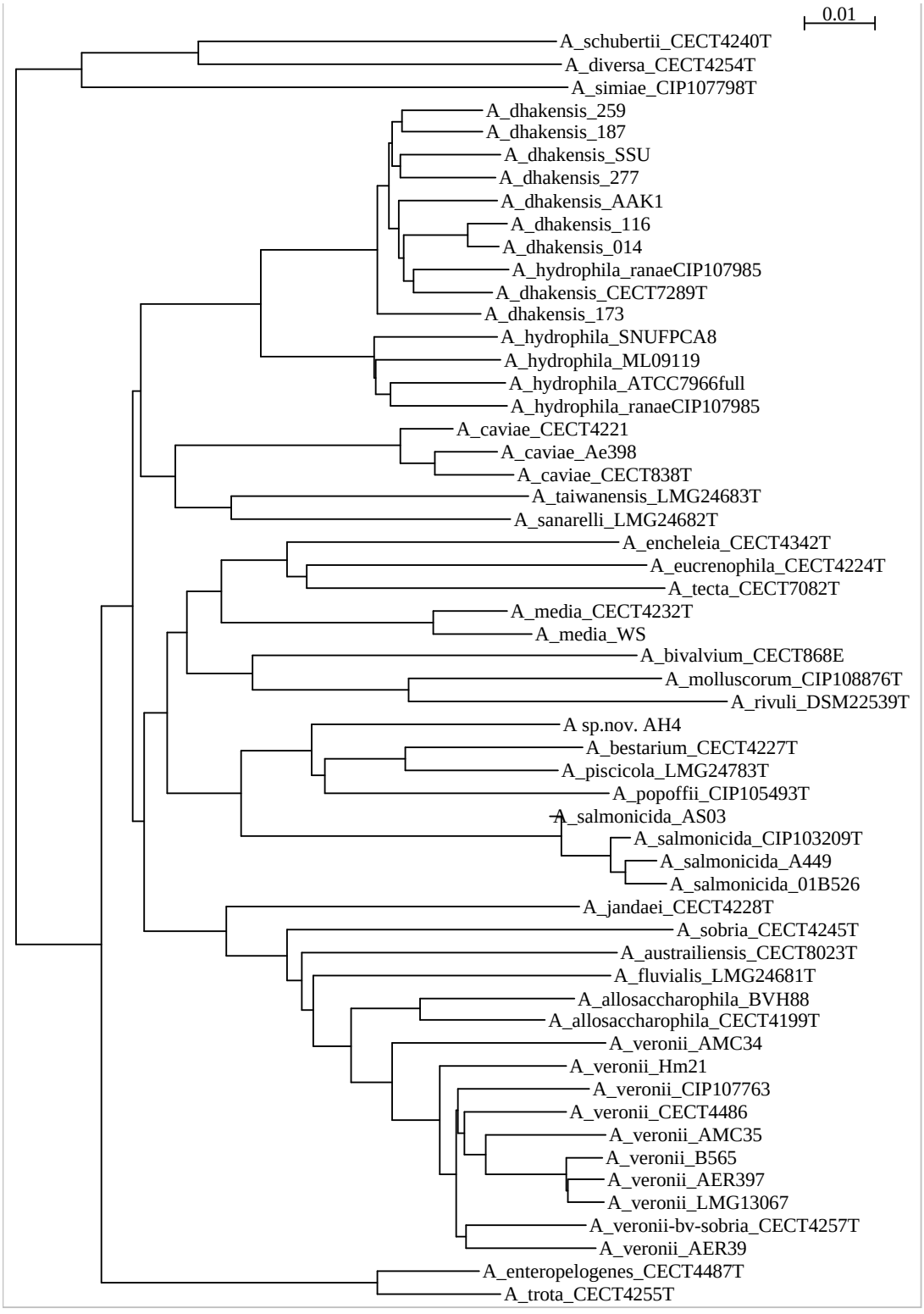


FigureS03. FastME phylogeny of the Roseo dataset inferred from jSpecies ANI values, converted to distances. Color-coding matches that used in Figure2ab and represents known clades of taxa. This method does not reconstruct accurate or even remotely sensible trees.

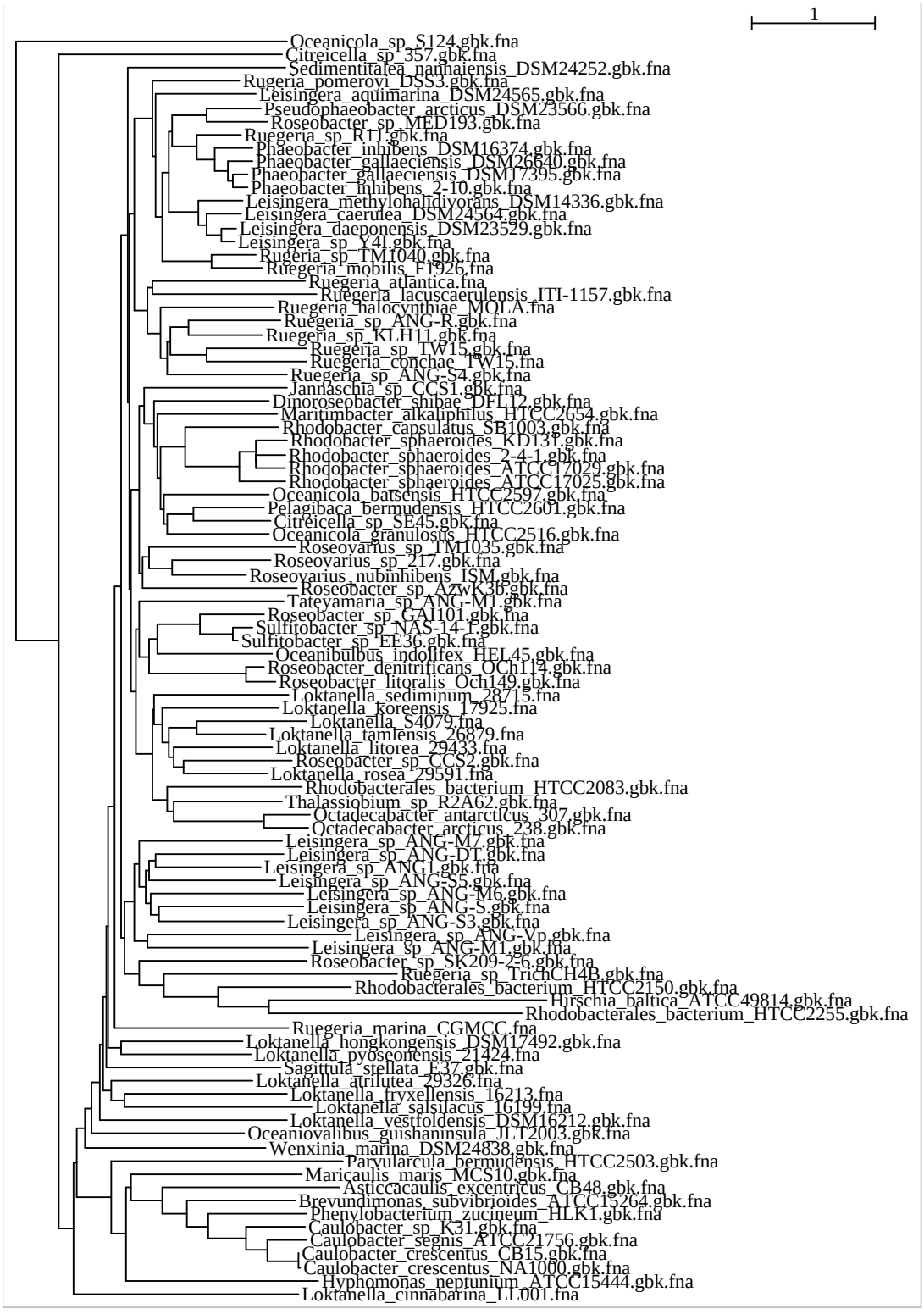


FigureS04. ANI-distance matrix of the AeroOG set. Highlighted are the taxa relationships with suspiciously large values. These values suggest these taxa do not belong in the Aeromonadaceae.

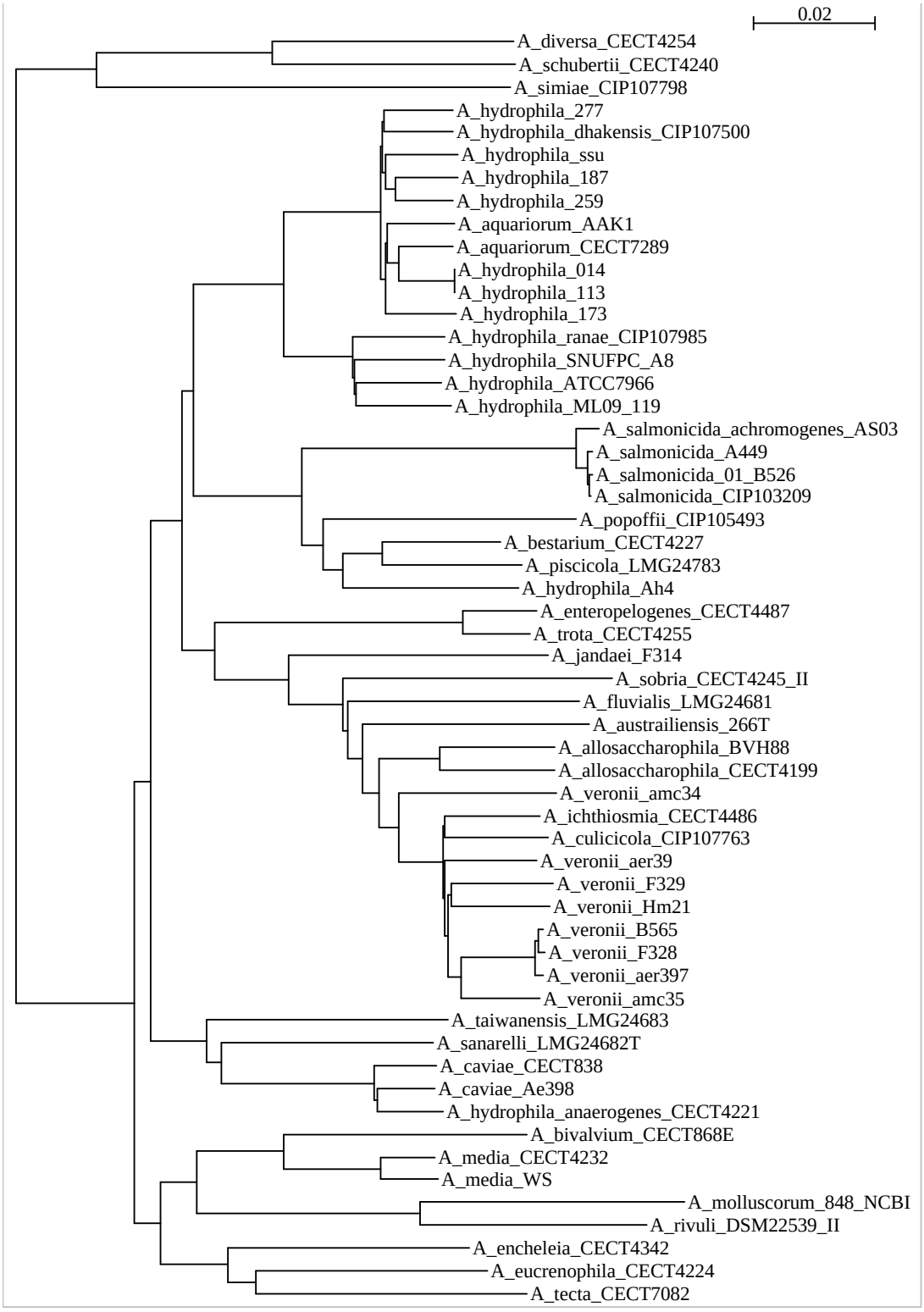
A.


0.01
A_schubertii_CECT4240T
A_diversa_CECT4254T
A_simiae_CIP107798T
A_dhakensis_259
A_dhakensis_187
A_dhakensis_SSU
A_dhakensis_277
A_dhakensis_AAK1
A_dhakensis_116
A_dhakensis_014
A_hydrophila_ranaeCIP107985
A_dhakensis_CECT7289T
A_dhakensis_173
A_hydrophila_SNUFPCA8
A_hydrophila_ML09119
A_hydrophila_ATCC7966full
A_hydrophila_ranaeCIP107985
A_caviae_CECT4221
A_caviae_Ae398
A_caviae_CECT838T
A_taiwanensis_LMG24683T
A_sanarelli_LMG24682T
A_encheleia_CECT4342T
A_eucrenophila_CECT4224T
A_tecta_CECT7082T
A_media_CECT4232T
A_media_WS
A_bivalvium_CECT868E
A_molluscorum_CIP108876T
A_rivuli_DSM22539T
A sp.nov. AH4
A_bestarium_CECT4227T
A_piscicola_LMG24783T
A_popoffii_CIP105493T
A_salmonicida_AS03
A_salmonicida_CIP103209T
A_salmonicida_A449
A_salmonicida_01B526
A_jandaei_CECT4228T
A_sobria_CECT4245T
A_austrailiensis_CECT8023T
A_fluvialis_LMG24681T
A_allosaccharophila_BVH88
A_allosaccharophila_CECT4199T
A_veronii_AMC34
A_veronii_Hm21
A_veronii_CIP107763
A_veronii_CECT4486
A_veronii_AMC35
A_veronii_B565
A_veronii_AER397
A_veronii_LMG13067
A_veronii-bv-sobria_CECT4257T
A_veronii_AER39
A_enteropelogenes_CECT4487T
A_trota_CECT4255T

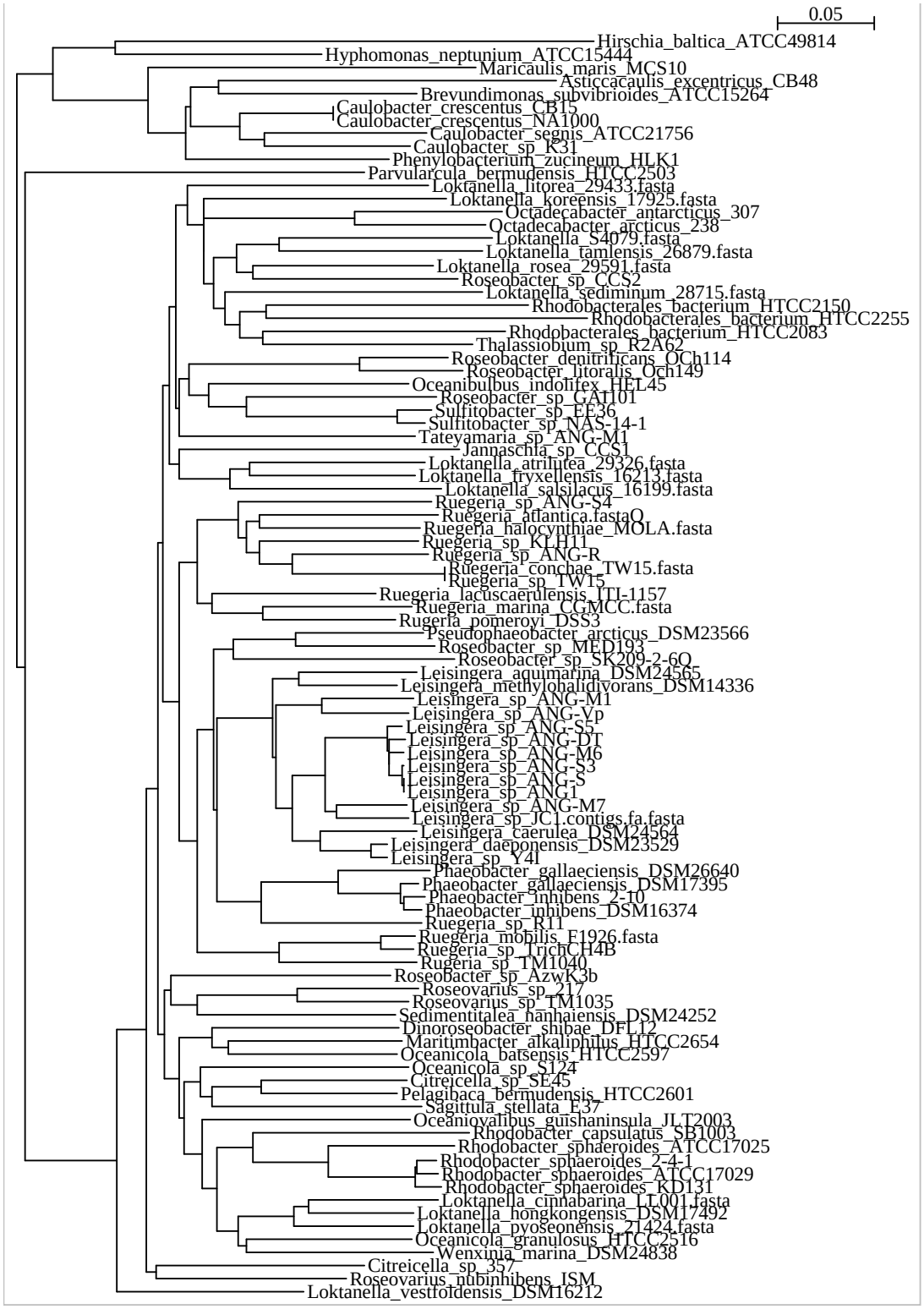
B.


1
Oceanicola_sp_S124.gbk.fna
Citreicella_sp_357.gbk.fna
Sedimentitalea_nanhaiensis_DSM24252.gbk.fna
Rugeria_pomeroyi_DSS3.gbk.fna
Leisingera_aquimarina_DSM24565.gbk.fna
Pseudophaeobacter_arcticus_DSM23566.gbk.fna
Roseobacter_sp_MED193.gbk.fna
Ruegeria_sp_R11.gbk.fna
Phaeobacter_inhibens_DSM16374.gbk.fna
Phaeobacter_gallaeciensis_DSM26640.gbk.fna
Phaeobacter_gallaeciensis_DSM17395.gbk.fna
Phaeobacter_inhibens_2-10.gbk.fna
Leisingera_methylohalidivorans_DSM14336.gbk.fna
Leisingera_caerulea_DSM24564.gbk.fna
Leisingera_daeponensis_DSM23529.gbk.fna
Leisingera_sp_Y4I.gbk.fna
Rugeria_sp_TM1040.gbk.fna
Ruegeria_mobilis_F1926.fna
Ruegeria_atlantica.fna
Ruegeria_lacuscaerulensis_ITI-1157.gbk.fna
Ruegeria_halocynthiae_MOLA.fna
Ruegeria_sp_ANG-R.gbk.fna
Ruegeria_sp_KLH11.gbk.fna
Ruegeria_sp_TW15.gbk.fna
Ruegeria_conchae_TW15.fna
Ruegeria_sp_ANG-S4.gbk.fna
Jannaschia_sp_CCS1.gbk.fna
Dinoroseobacter_shibae_DFL12.gbk.fna
Maritimbacter_alkaliphilus_HTCC2654.gbk.fna
Rhodobacter_capsulatus_SB1003.gbk.fna
Rhodobacter_sphaeroides_KD131.gbk.fna
Rhodobacter_sphaeroides_2.4.1.gbk.fna
Rhodobacter_sphaeroides_ATCC17029.gbk.fna
Rhodobacter_sphaeroides_ATCC17025.gbk.fna
Oceanicola_batsensis_HTCC2597.gbk.fna
Pelagibaca_bermudensis_HTCC2601.gbk.fna
Citreicella_sp_SE45.gbk.fna
Oceanicola_granulosus_HTCC2516.gbk.fna
Roseovarius_sp_TM1035.gbk.fna
Roseovarius_sp_217.gbk.fna
Roseovarius_nubinhibens_ISM.gbk.fna
Roseobacter_sp_AzwK3b.gbk.fna
Tateyamaria_sp_ANG-M1.gbk.fna
Roseobacter_sp_GAI101.gbk.fna
Sulfitobacter_sp_NAS-14.1.gbk.fna
Sulfitobacter_sp_EE36.gbk.fna
Oceanibulbus_indolifex_HEL45.gbk.fna
Roseobacter_denitrificans_OCh114.gbk.fna
Roseobacter_litoralis_Och149.gbk.fna
Loktanella_sediminum_28715.fna
Loktanella_koreensis_17925.fna
Loktanella_S4079.fna
Loktanella_tamlensis_26879.fna
Loktanella_litorea_29433.fna
Roseobacter_sp_CCS2.gbk.fna
Loktanella_rosea_29591.fna
Rhodobacterales_bacterium_HTCC2083.gbk.fna
Thalassiobium_sp_R2A62.gbk.fna
Octadecabacter_antarcticus_307.gbk.fna
Octadecabacter_arcticus_238.gbk.fna
Leisingera_sp_ANG-M7.gbk.fna
Leisingera_sp_ANG-DT.gbk.fna
Leisingera_sp_ANG1.gbk.fna
Leisingera_sp_ANG-S5.gbk.fna
Leisingera_sp_ANG-M6.gbk.fna
Leisingera_sp_ANG-S.gbk.fna
Leisingera_sp_ANG-S3.gbk.fna
Leisingera_sp_ANG-Vp.gbk.fna
Leisingera_sp_ANG-M1.gbk.fna
Roseobacter_sp_SK209-2-6.gbk.fna
Ruegeria_sp_TrichCH4B.gbk.fna
Rhodobacterales_bacterium_HTCC2150.gbk.fna
Hirschia_baltica_ATCC49814.gbk.fna
Rhodobacterales_bacterium_HTCC2255.gbk.fna
Ruegeria_marina_CGMCC.fna
Loktanella_hongkongensis_DSM17492.gbk.fna
Loktanella_pyoseonensis_21424.fna
Sagittula_stellata_E37.gbk.fna
Loktanella_atrilutea_29326.fna
Loktanella_fryxellensis_16213.fna
Loktanella_salsilacus_16199.fna
Loktanella_vestfoldensis_DSM16212.gbk.fna
Oceaniovalibus_guishaninsula_JLT2003.gbk.fna
Wenxinia_marina_DSM24838.gbk.fna
Parvularcula_bermudensis_HTCC2503.gbk.fna
Maricaulis_maris_MCS10.gbk.fna
Asticcacaulis_excentricus_CB48.gbk.fna
Brevundimonas_subvibrioides_ATCC15264.gbk.fna
Phenylobacterium_zucineum_HLK1.gbk.fna
Caulobacter_sp_K31.gbk.fna
Caulobacter_segnis_ATCC21756.gbk.fna
Caulobacter_crescentus_CB15.gbk.fna
Caulobacter_crescentus_NA1000.gbk.fna
Hyphomonas_neptunium_ATCC15444.gbk.fna
Loktanella_cinnabarina_LL001.fna

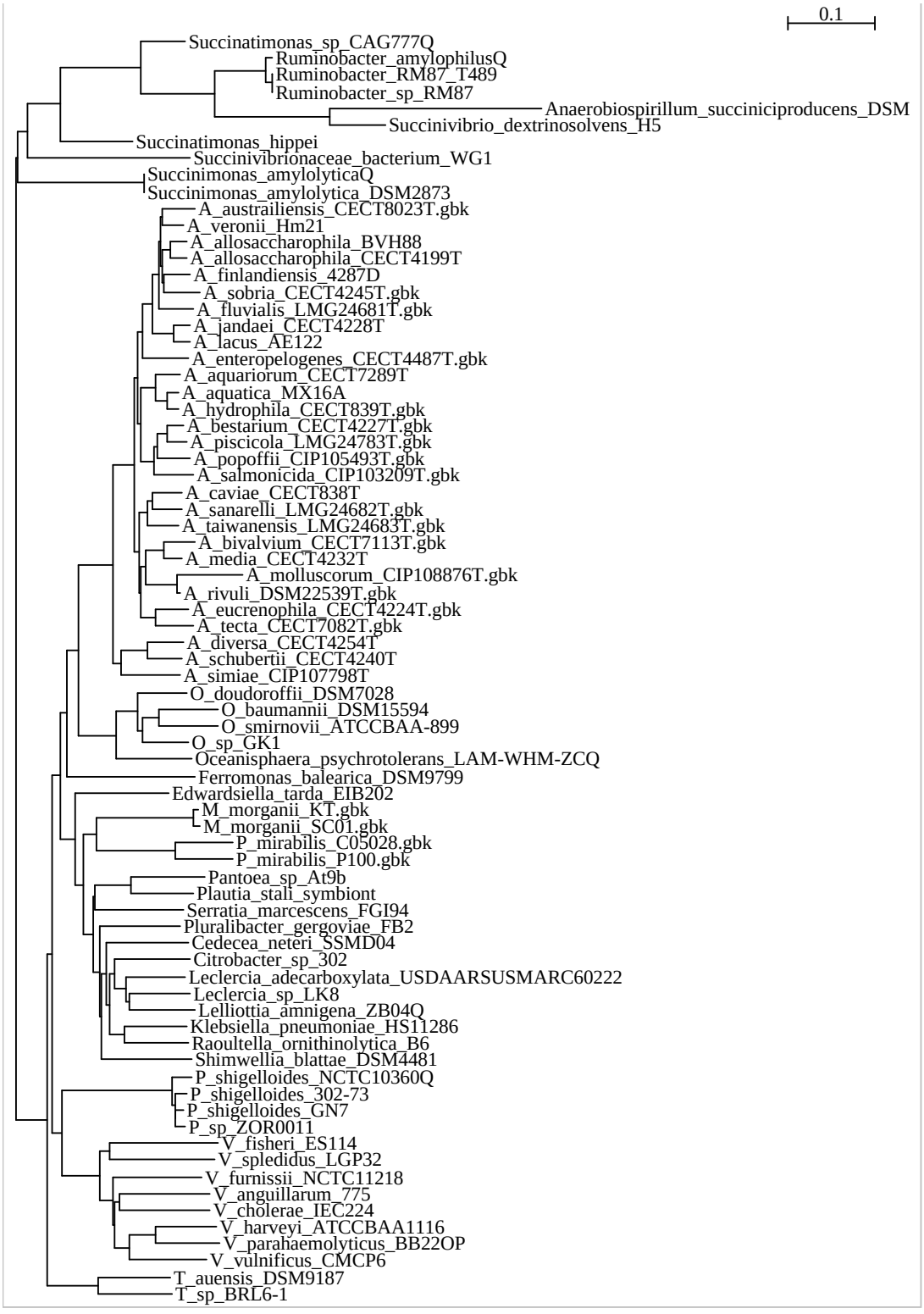
C.


0.02
A_diversa_CECT4254
A_schubertii_CECT4240
A_simiae_CIP107798
A_hydrophila_277
A_hydrophila_dhakensis_CIP107500
A_hydrophila_ssu
A_hydrophila_187
A_hydrophila_259
A_aquariorum_AAK1
A_aquariorum_CECT7289
A_hydrophila_014
A_hydrophila_113
A_hydrophila_173
A_hydrophila_ranae_CIP107985
A_hydrophila_SNUFPC_A8
A_hydrophila_ATCC7966
A_hydrophila_ML09_119
A_salmonicida_achromogenes_AS03
A_salmonicida_A449
A_salmonicida_01_B526
A_salmonicida_CIP103209
A_popoffii_CIP105493
A_bestarium_CECT4227
A_piscicola_LMG24783
A_hydrophila_Ah4
A_enteropelogenes_CECT4487
A_trota_CECT4255
A_jandaei_F314
A_sobria_CECT4245_II
A_fluvialis_LMG24681
A_austrailiensis_266T
A_allosaccharophila_BVH88
A_allosaccharophila_CECT4199
A_veronii_amc34
A_ichthiosmia_CECT4486
A_culicicola_CIP107763
A_veronii_aer39
A_veronii_F329
A_veronii_Hm21
A_veronii_B565
A_veronii_F328
A_veronii_aer397
A_veronii_amc35
A_taiwanensis_LMG24683
A_sanarelli_LMG24682T
A_caviae_CECT838
A_caviae_Ae398
A_hydrophila_anaerogenes_CECT4221
A_bivalvium_CECT868E
A_media_CECT4232
A_media_WS
A_molluscorum_848_NCBI
A_rivuli_DSM22539_II
A_encheleia_CECT4342
A_eucrenophila_CECT4224
A_tecta_CECT7082

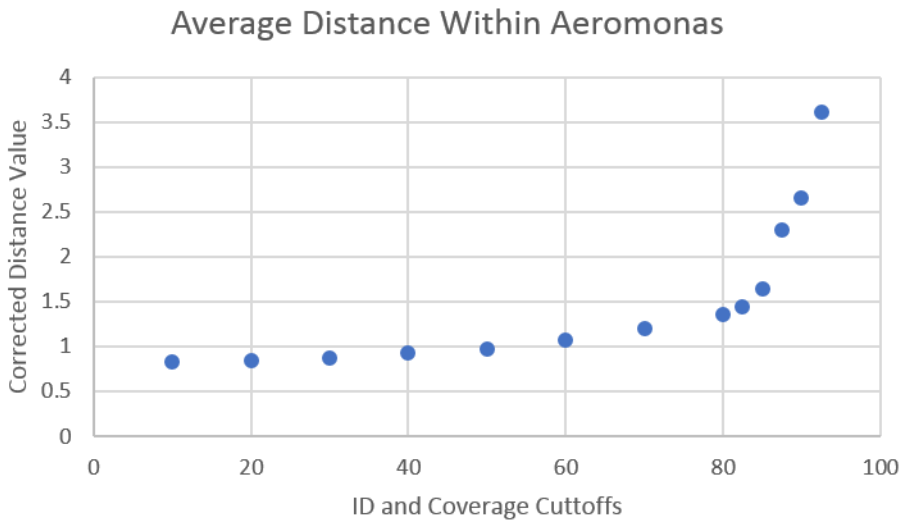
D.


0.05
Hirschia_baltica_ATCC49814
Hyphomonas_neptunium_ATCC15444
Maricaulis_maris_MCS10
Asticcacaulis_excentricus_CB48
Brevundimonas_subvibrioides_ATCC15264
Caulobacter_crescentus_CB15
Caulobacter_crescentus_NA1000
Caulobacter_segnis_ATCC21756
Caulobacter_sp_K31
Phenylobacterium_zucineum_HLK1
Parvularcula_bermudensis_HTCC2503
Loktanella_litorea_29433.fasta
Loktanella_koreensis_17925.fasta
Octadecabacter_antarcticus_307
Octadecabacter_arcticus_238
Loktanella_S4079.fasta
Loktanella_tamlensis_26879.fasta
Loktanella_rosea_29591.fasta
Roseobacter_sp_CCS2
Loktanella_sediminum_28715.fasta
Rhodobacterales_bacterium_HTCC2150
Rhodobacterales_bacterium_HTCC2255
Rhodobacterales_bacterium_HTCC2083
Thalassiobium_sp_R2A62
Roseobacter_denitrificans_OCh114
Roseobacter_litoralis_Och149
Oceanibulbus_indolifex_HEL45
Roseobacter_sp_GAI101
Sulfitobacter_sp_EE36
Sulfitobacter_sp_NAS-14-1
Tateyamaria_sp_ANG-M1
Jannaschia_sp_CCS1
Loktanella_atrilutea_29326.fasta
Loktanella_fryxellensis_16213.fasta
Loktanella_salsilacus_16199.fasta
Ruegeria_sp_ANG-S4
Ruegeria_atlantica.fastaO
Ruegeria_halocynthiae_MOLA.fasta
Ruegeria_sp_KLH11
Ruegeria_sp_ANG-R
Ruegeria_conchae_TW15.fasta
Ruegeria_sp_TW15
Ruegeria_lacuscaerulensis_ITI-1157
Ruegeria_marina_CGMCC.fasta
Rugeria_pomeroyi_DSS3
Pseudophaeobacter_arcticus_DSM23566
Roseobacter_sp_MED193
Roseobacter_sp_SK209-2-6O
Leisingera_aquimarina_DSM24565
Leisingera_methylohalidivorans_DSM14336
Leisingera_sp_ANG-M1
Leisingera_sp_ANG-Vp
Leisingera_sp_ANG-S5
Leisingera_sp_ANG-DT
Leisingera_sp_ANG-M6
Leisingera_sp_ANG-S3
Leisingera_sp_ANG-S
Leisingera_sp_ANG1
Leisingera_sp_ANG-M7
Leisingera_sp_JC1.contigs.fa.fasta
Leisingera_caerulea_DSM24564
Leisingera_daeponensis_DSM23529
Leisingera_sp_Y4I
Phaeobacter_gallaeciensis_DSM26640
Phaeobacter_gallaeciensis_DSM17395
Phaeobacter_inhibens_2-10
Phaeobacter_inhibens_DSM16374
Ruegeria_sp_R11
Ruegeria_mobilis_F1926.fasta
Ruegeria_sp_TrichCH4B
Rugeria_sp_TM1040
Roseobacter_sp_AzwK3b
Roseovarius_sp_217
Roseovarius_sp_TM1035
Sedimentitalea_nanhaiensis_DSM24252
Dinoroseobacter_shibae_DFL12
Maritimbacter_alkaliphilus_HTCC2654
Oceanicola_batsensis_HTCC2597
Oceanicola_sp_S124
Citreicella_sp_SE45
Pelagibaca_bermudensis_HTCC2601
Sagittula_stellata_E37
Oceaniovalibus_guishaninsula_JLT2003
Rhodobacter_capsulatus_SB1003
Rhodobacter_sphaeroides_ATCC17025
Rhodobacter_sphaeroides_2-4-1
Rhodobacter_sphaeroides_ATCC17029
Rhodobacter_sphaeroides_KD131
Loktanella_cinnabarina_LL001.fasta
Loktanella_hongkongensis_DSM17492
Loktanella_pyoseonensis_21424.fasta
Oceanicola_granulosus_HTCC2516
Wenxinia_marina_DSM24838
Citreicella_sp_357
Roseovarius_nubinhibens_ISM
Loktanella_vestfoldensis_DSM16212

E.


0.1
Succinatimonas_sp_CAG777Q
Ruminobacter_amylophilusQ
Ruminobacter_RM87_T489
Ruminobacter_sp_RM87
Anaerobiospirillum_succiniciproducens_DSM
Succinivibrio_dextrinosolvens_H5
Succinatimonas_hippei
Succinivibrionaceae_bacterium_WG1
Succinimonas_amylolyticaQ
Succinimonas_amylolytica_DSM2873
A_austrailiensis_CECT8023T.gbk
A_veronii_Hm21
A_allosaccharophila_BVH88
A_allosaccharophila_CECT4199T
A_finlandiensis_4287D
A_sobria_CECT4245T.gbk
A_fluvialis_LMG24681T.gbk
A_jandaei_CECT4228T
A_lacus_AE122
A_enteropelogenes_CECT4487T.gbk
A_aquariorum_CECT7289T
A_aquatica_MX16A
A_hydrophila_CECT839T.gbk
A_bestarium_CECT4227T.gbk
A_piscicola_LMG24783T.gbk
A_popoffii_CIP105493T.gbk
A_salmonicida_CIP103209T.gbk
A_caviae_CECT838T
A_sanarelli_LMG24682T.gbk
A_taiwanensis_LMG24683T.gbk
A_bivalvium_CECT7113T.gbk
A_media_CECT4232T
A_molluscorum_CIP108876T.gbk
A_rivuli_DSM22539T.gbk
A_eucrenophila_CECT4224T.gbk
A_tecta_CECT7082T.gbk
A_diversa_CECT4254T
A_schubertii_CECT4240T
A_simiae_CIP107798T
O_doudoroffii_DSM7028
O_baumannii_DSM15594
O_smirnovii_ATCCBAA-899
O_sp_GK1
Oceanisphaera_psychrotolerans_LAM-WHM-ZCQ
Ferromonas_balearica_DSM9799
Edwardsiella_tarda_EIB202
M_morganii_KT.gbk
M_morganii_SC01.gbk
P_mirabilis_C05028.gbk
P_mirabilis_P100.gbk
Pantoea_sp_At9b
Plautia_stali_symbiont
Serratia_marcescens_FGI94
Pluralibacter_gergoviae_FB2
Cedecea_neteri_SSMD04
Citrobacter_sp_302
Leclercia_adecarboxylata_USDAARSUSMARC60222
Leclercia_sp_LK8
Lelliottia_amnigena_ZB04Q
Klebsiella_pneumoniae_HS11286
Raoultella_ornithinolytica_B6
Shimwellia_blattae_DSM4481
P_shigelloides_NCTC10360Q
P_shigelloides_302-73
P_shigelloides_GN7
P_sp_ZOR0011
V_fisheri_ES114
V_splendidus_LGP32
V_furnissii_NCTC11218
V_anguillarum_775
V_cholerae_IEC224
V_harveyi_ATCCBAA1116
V_parahaemolyticus_BB22OP
V_vulnificus_CMCP6
T_auensis_DSM9187
T_sp_BRL6-1

FigureS05. Phylogenies derived from other whole-genome phylogeny methods. Panels A & B: mBio & Roseo dataset trees created with GBDP. Panels C,D,E: mBio, Roseo, and AeroOG datasets created with Mashtree.



FigureS06. The average distance value from within *Aeromonas* comparisons with varying cutoffs for the %ID and Coverage percentage needed to include a match in the calculation.

TableS01. List of reclassifications proposed as well as the categorization used for ROC taxonomic discrimination analyses.

Changes for Species Level Cuttoff Comparison

Current ID	Species ID
Leisingera **daeponensis DSM23529**	**Leis_2**
Leisingera **sp ANG1**	**Leis_1**
Leisingera **sp ANG-M6**	**Leis_1**
Leisingera **sp ANG-S**	**Leis_1**
Leisingera **sp ANG-S3**	**Leis_1**
Leisingera **sp ANG-S5**	**Leis_1**
Leisingera **sp Y4I**	**Leis_2**
Leisingera **sp ANG-DT**	**Leis_1**
Sulfitobacter **sp EE36**	**Sulf_1**
Sulfitobacter **sp NAS-14-1**	**Sulf_1**
Ruegeria **conchae TW15**	**Rue_conchae**
Ruegeria **lacuscaerulensis ITI-1157**	**Rue_lacus**

***Ruegeria* mobilis F1926**	**Rue_mobilis**
***Ruegeria* sp TrichCH4B**	**Rue_mobilis**
***Ruegeria* sp TW15**	**Rue_conchae**
***Silicibacter* lacuscaerulensis ITI1157**	**Rue_lacus**

Suggested Reclassifactions

Note: These suggested changes were applied to classifications in the higher order ROC comparisons

Name	Evidence	Reccomendation
***Rugeria* sp TM1040** ***Ruegeria* mobilis F1926** ***Ruegeria* sp TrichCH4B**	Previou studies grouping these taxa with Ruegeria lacked resolution, and had poor branch support. We find this group consistently seperating from the rest of the genus despite changing methodologies.	Reclassify as a separate genus with the following members. Additionally, Reclassify TrichCH4B as a mobilis strain.
***Loktanella* cinnabarina LL001** ***Loktanella* hongkongensis DSM17492** ***Loktanella* pyoseonensis 21424**	Similar reasons to those stated for the *Ruegeria* group. See the main text for citations.	Reclassify as a separate genus with these members.
***Ruegeria* sp R11**	Consistently groups as an outgroup to Phaeobacter no matter the methodology	Investigate relationship to Phaeobacter, and closely related taxa. Certainly should not be a member of *Ruegeria*
Anaerobiospirillum succiniciproducens DSM **Oceanisphaera psychrotolerans LAM-WHM-ZC** **Ruminobacter amylophilus DSM 1361** **Ruminobacter RM87 T489** **Succinatimonas hippei YIT 12066** **Succinatimonas sp CAG777**	The classification of this grouping as part of Aeromonadales is based off a single 1999 study by Hippe et al. Within said study, there is a single 16s tree, from which all classification descisions regarding this group is derived. However,even within that study, branch support is less than 50 for the node connecting this group to *Aeromonas,* and within the group o *Aeromonas,* there are members of gammaproteobacteria that are not a part of the *Aeromonadales,* yet subsequent authors have used this phylogeny to claim the Succinivibrionaceae should be part of the Aeromonadales.	They should certainly not be a part of the Aeromonadales, and likely deserve a separate order designation. However, the family level classifications within this order are not clear, as these taxa sit on long branches. Additionaly sampling from the gammaproteobacteria may help to break up these branches and provide a better understanding of the placement of families within this group.

Succinimonas amylolytica DSM2873
Succinivibrio dextrinosolvens H5
Succinivibrio Phil9
Succinivibrionaceae bacterium WG1

TableS02. List of genomes used in this study, organized by datasets.

Species Name	Strain	Genome Length (Mbp)	No. of Scaffolds	Asscesion Number
Aeromonas mBio				
A allosaccharophila	BVH88	4.71	131	NZ_CDCB000000000.1
A allosaccharophila	CECT4199	4.66	120	NZ_CDBR000000000.1
A austrailiensis	266T	4.11	113	NZ_CDDH000000000.1
A bestarium	CECT4227	4.69	41	NZ_CDDA000000000.1
A bivalvium	CECT868E	5.50	1112	NZ_CDBT000000000.1
A caviae	Ae398	4.44	149	NZ_CACP000000000.1
A caviae	CECT838	4.47	111	NZ_CDBK000000000.1
A caviae	CECT4221	4.58	332	NZ_CDBS000000000.1
A dhakensis	AAK1	4.76	36	NZ_BAFL000000000.1
A dhakensis	CECT7289	4.69	78	NZ_CDBP000000000.1
A dhakensis	116	4.68	45	NZ_ANPN000000000.1
A dhakensis	187	4.78	59	NZ_AOBO000000000.1
A dhakensis	CIP107500	4.71	73	NZ_CDBH000000000.1
A diversa	CECT4254	4.06	37	NZ_CDCE000000000.1
A encheleia	CECT4342	4.47	35	NZ_CDDI000000000.1
A enteropelogenes	CECT4487	4.47	46	NZ_CDCG000000000.1
A enteropelogenes	CECT4255	4.34	27	NZ_CDDE000000000.1
A eucrenophila	CECT4224	4.54	22	NZ_CDDF000000000.1
A fluvialis	LMG24681	3.90	76	NZ_CDBO000000000.1
A hydrophila	14	4.67	75	NZ_AOBM000000000.1
A hydrophila	173	4.79	74	NZ_AOBN000000000.1
A hydrophila	259	4.70	80	NZ_AOBP000000000.1
A hydrophila	277	4.79	41	NZ_AOBQ000000000.1
A hydrophila	ATCC7966	4.74	1	NC_0085700.1
A hydrophila	ML09 119	5.02	1	NC_0212900.1
A hydrophila	SNUFPC-A8	4.97	41	NZ_AMQA000000000.1
A hydrophila	ssu	4.94	2	NZ_AGWR000000000.1
A hydrophila subsp. *ranae*	CIP107985	4.68	107	NZ_CDDC000000000.1

A jandaei	CECT 4228	4.50	58	NZ_CDBV000000000.1
A media	CECT4232	4.48	233	NZ_CDBZ000000000.1
A media	WS	4.32	258	NZ_CP0075670.1
A molluscorum	848	4.24	309	NZ_AQGQ000000000.1
A piscicola	LMG 24783	5.18	91	NZ_CDBL000000000.1
A popoffii	CIP105493	4.76	105	NZ_CDBI000000000.1
A rivuli	DSM22539 II	4.53	102	NZ_CDBJ000000000.1
A salmonicida	01 B526	4.93	604	NZ_AGVO0000000000.1
A salmonicida	A449	5.04	6	NC_0093480.1
A salmonicida	CIP103209	4.74	128	NZ_CDDW000000000.1
*A salmonicida*subsp. a*chromogenes*	AS03	4.45	342	NZ_AMQG000000000.1
A sanarelli	LMG24682T	4.19	98	NZ_CDBN000000000.1
A schubertii	CECT4240	4.13	111	NZ_CDDB000000000.1
A simiae	CIP107798	3.99	100	NZ_CDBY000000000.1
A sobria	CECT4245	4.68	48	NZ_CDBW000000000.1
A sp. nov	Ah4	4.87	41	SAMEA2752429
A sp. nov	amc34	4.58	1	NZ_AGWU000000000.1
A taiwanensis	LMG24683	5.08	987	NZ_CDDD000000000.1
A tecta	CECT7082	4.76	51	NZ_CDCA000000000.1
A veronii	CIP107763	4.43	64	NZ_CDDU000000000.1
A veronii	CECT4486	4.41	66	NZ_CDBU000000000.1
A veronii	aer39	4.42	4	NZ_AGWT000000000.1
A veronii	aer397	4.50	5	NZ_AGWV000000000.1
A veronii	amc35	4.57	2	NZ_AGWW000000000.1
A veronii	B565	4.55	1	NC_0154240.1
A veronii	F328	4.52	52	NZ_CDDK000000000.1
A veronii	Hm21	4.68	50	NZ_ATFB000000000.1

Species Name	**Strain**	**Genome Length (Mbp)**	**No. of Scaffolds**	**Asscesion Number**
Aeromonas AeroOG				
A allosaccharophila	BVH88	4.71	131	NZ_CDCB000000000.1
A allosaccharophila	CECT4199	4.66	120	NZ_CDBR000000000.1
A aquariorum	CECT7289T	4.69	78	NZ_CDBP00000000.1
A aquatica	MX16A	4.78	1	NZ_CP018201.1
A austrailiensis	266T	4.11	113	NZ_CDDH000000000.1
A bestarium	CECT4227	4.69	41	NZ_CDDA000000000.1
A bivalvium	CECT868E	5.50	1112	NZ_CDBT000000000.1
A cavernicola	642.176	3.92	341	NZ_PGGC00000000.1
A caviae	CECT838T	4.47	111	NZ_CDBK000000000.1
A diversa	CECT4254	4.06	37	NZ_CDCE000000000.1
A enteropelogenes	CECT4487	4.47	46	NZ_CDCG000000000.1

A eucrenophila	CECT4224	4.54	22	NZ_CDDF000000000.1
A finlandiensis	4287D	4.72	376	JRGK01000001.1
A fluvialis	LMG24681	3.90	76	NZ_CDBO000000000.1
A hydrophila	CECT839T	4.74	1	NC_0085700.1
A jandaei	CECT 4228	4.50	58	NZ_CDBV000000000.1
A lacus	AE122	4.39	196	JRGM01000001.1
A lusitana	642.175	4.55	67	PGCP00000000.1
A media	CECT4232	4.48	233	NZ_CDBZ000000000.1
A molluscorum	CIP108876T	4.24	309	NZ_AQGQ000000000.1
A piscicola	LMG 24783	5.18	91	NZ_CDBL000000000.1
A popoffii	CIP105493	4.76	105	NZ_CDBI000000000.1
A rivuli	DSM22539 II	4.53	102	NZ_CDBJ000000000.1
A salmonicida	CIP103209	4.74	128	NZ_CDDW000000000.1
A sanarelli	LMG24682T	4.19	98	NZ_CDBN000000000.1
A schubertii	CECT4240	4.13	111	NZ_CDDB000000000.1
A simiae	CIP107798	3.99	100	NZ_CDBY000000000.1
A sobria	CECT4245	4.68	48	NZ_CDBW000000000.1
A taiwanensis	LMG24683	5.08	987	NZ_CDDD000000000.1
A tecta	CECT7082	4.76	51	NZ_CDCA000000000.1
A veronii	Hm21	4.68	50	NZ_ATFB000000000.1
Anaerobiospirillum succiniciproducens	DSM	3.80	155	AXWV01000001.1
Cedecea neteri	SSMD04	4.88	1	NZ_CP009451.1
Citrobacter	sp 302	5.02	9	NZ_KI391984.1
Edwardsiella tarda	EIB202	3.80	2	NC_013508.1
Ferromonas balearica	DSM9799	4.28	1	NC_014541.1
Klebsiella pneumoniae	HS11286	5.68	7	NC_016845.1
Leclercia adecarboxylata	USDAARSUSMARC60222	4.80	1	NZ_CP013990.1
Leclercia	sp LK8	5.21	75	NZ_LDUO01000001.1
Lelliottia amnigena	ZB04	4.62	1	NZ_CP015774.1
M morganii	KT	3.83	58	NC_020418.1
M morganii	SC01	4.15	63	NZ_AMWL00000000.2
Oceanimonas baumannii	DSM15594s	3.75	31	2593339295*
Oceanimonas doudoroffii	DSM7028s	3.83	194	2506520053*
Oceanimonas smirnovii	ATCCBAA-899s	3.28	28	NZ_ARMW00000000.1
Oceanimonas	sp GK1s	3.51	1	NC_016745.1
Oceanisphaera psychrotolerans	LAM-WHM-ZC	3.82	70	MDKE01000001.1
Plesiomonas shigelloides	302-73s	3.91	389	NZ_AQQO00000000.1
Plesiomonas shigelloides	GN7s	3.92	83	NZ_JWHQ00000000.1
Plesiomonas shigelloides	NCTC10360s	2.46	1	NZ_LT575468.1
Plesiomonas	sp ZOR0011s	3.84	152	NZ_JRKB00000000.1
Proteus mirabilis	C05028	3.79	85	NZ_ANBT00000000.1
Proteus mirabilis	P100	4.15	126	2740892266*
Pantoea	sp At9b	6.31	6	NC_014837.1
Plautia	stali symbiont	4.09	3	NC_022546.1
Pluralibacter gergoviae	FB2	5.49	1	NZ_CP009450.1
Raoultella ornithinolytica	B6	5.40	1	NC_021066.1

Ruminobacter amylophilus	DSM 1361	2.82	117	FOXF01000116.1
Ruminobacter	RM87 T489	2.86	123	JNKD01000001.1
Ruminobacter	sp RM87	2.86	123	NZ_JNKD01000001.1
Serratia marcescens	FGI94	4.86	1	NC_020064.1
Shimwellia blattae	DSM4481	4.16	1	NC_017910.1
Succinimonas amylolytica	DSM 2873	3.96	23	KB899636.1
Succinatimonas hippei	YIT 12066	2.31	141	GL830939.1
Succinatimonas	sp CAG777	2.25	71	HF987897.1
Succinivibrionaceae bacterium	WG1	2.95	43	GL995195.1
Succinimonas amylolytica	DSM2873	3.96	23	NZ_KB899636.1
Succinivibrio dextrinosolvens	H5	2.68	106	NZ_KL370853.1
Tolumonas auensis	DSM9187s	3.47	1	NC_012691.1
Tolumonas	sp BRL6-1s	3.63	9	NZ_AZUK00000000.1
Vibrio anguillarum	775s	4.05	2	NC_015633.1
Vibrio cholerae	IEC224s	4.08	2	NC_016944.1
Vibrio fisheri	ES114s	4.27	3	NC_006840.2
Vibrio furnissii	NCTC11218s	4.92	2	NC_016602.1
Vibrio harveyi	ATCCBAA1116s	6.06	3	NC_009777.1
Vibrio parahaemolyticus	BB22OPs	5.10	2	NC_019955.1
Vibrio splendidus	LGP32s	4.97	2	NC_011744.2
Vibrio vulnificus	CMCP6s	5.13	2	NC_004459.3

Species Name	**Strain**	**Genome Length (Mbp)**	**No. of Scaffolds**	**Asscesion Number**
Roseobacter Set				
Asticcacaulis excentricus	CB48	4.31	4	NC_014817.1
Brevundimonas subvibrioides	ATCC15264	3.45	1	NC_014375.1
Caulobacter	K31	5.89	3	NC_002696.2
Caulobacter crescentus	CB15	4.02	1	NC_011916.1
Caulobacter crescentus	NA1000	4.04	1	NC_014100.1
Caulobacter segnis	ATCC21756	4.66	1	NC_010333.1
Citreicella	357	4.60	180	NZ_AJKJ01000164.1
Citreicella	SE45	5.52	9	NZ_GG704601.1
Dinoroseobacter shibae	DFL12	4.42	6	NC_009959.1
Hirschia baltica	ATCC49814	3.54	2	NC_012983.1
Hyphomonas neptunium	ATCC15444	3.71	1	NC_008358.1
Jannaschia	CCS1	4.40	2	NC_007802.1
Leisingera	ANG-DT	4.60	91	NZ_JWLE00000000.1
Leisingera	ANG-M1	5.38	92	NZ_JWLC00000000.1
Leisingera	ANG-M6	4.54	54	NZ_JWLG00000000.1
Leisingera	ANG-M7	4.58	58	NC_023146.1
Leisingera	ANG-S	4.57	68	NZ_JWLM00000000.1
Leisingera	ANG-S3	4.60	70	NZ_JWLF00000000.1
Leisingera	ANG-S5	4.66	43	NZ_JWLH00000000.1

Leisingera	ANG-Vp	5.15	143	NZ_JWLD00000000.1
Leisingera	ANG1	4.60	26	NZ_AFCF00000000.2
Leisingera	Y4I	4.34	5	NZ_DS995283.1
Leisingera	JC1	5.19	168	NZ_LYUZ00000000.1
Leisingera aquimarina	DSM24565	5.34	15	NC_023146.1
Leisingera caerulea	DSM24564	5.34	21	NZ_AXBI00000000.1
Leisingera daeponensis	DSM23529	4.64	12	NZ_AXBD00000000.1
Leisingera methylohalidivorans	DSM14336	4.65	3	NZ_DS995283.1
Loktanella	S4079	3.56	43	NZ_JXYE00000000.1
Loktanella atrilutea	29326	4.21	46	NZ_FQUE00000000.1
Loktanella cinnabarina	LL001	3.90	192	NZ_BATB00000000.1
Loktanella fryxellensis	16213	3.55	75	FOCI00000000.1
Loktanella hongkongensis	DSM17492	3.19	16	NZ_KB823002.1
Loktanella koreensis	17925	3.65	4	FOIZ00000000.1
Loktanella litorea	29433	3.32	8	FOZM00000000.1
Loktanella pyoseonensis	21424	3.91	22	NZ_FTPR00000000.1
Loktanella rosea	29591	3.51	5	FNAT00000000.1
Loktanella salsilacus	16199	4.13	77	FOTF00000000.1
Loktanella sediminum	28715	3.26	16	NZ_FQXB00000000.1
Loktanella tamlensis	26879	3.19	9	FOYP00000000.1
Loktanella vestfoldensis	DSM16212	3.72	49	NZ_ARNL00000000.1
Maricaulis maris	MCS10	3.37	1	NC_008347.1
Maritimbacter alkaliphilus	HTCC2654	4.54	7	NZ_AAMT01000046.1
Oceanibulbus indolifex	HEL45	4.11	105	NZ_ABID01000017.1
Oceanicola	S124	4.65	339	NZ_AAMO01000007.1
Oceanicola batsensis	HTCC2597	4.44	7	NZ_CH724110.1
Oceanicola granulosus	HTCC2516	4.05	9	NZ_AFPM01000263.1
Oceaniovalibus guishaninsula	JLT2003	2.90	68	NZ_AMGO01000046.1
Octadecabacter antarcticus	307	4.88	2	NC_020911.1
Octadecabacter arcticus	238	5.48	3	NC_020909.1
Parvularcula bermudensis	HTCC2503	2.90	1	NC_014414.1
Pelagibaca bermudensis	HTCC2601	5.48	6	NZ_DS022279.1
Phaeobacter gallaeciensis	DSM17395	4.23	4	NC_018290.1
Phaeobacter gallaeciensis	DSM26640	4.54	8	NC_023143.1
Phaeobacter inhibens	2-10	4.16	4	NC_018423.1
Phaeobacter inhibens	DSM16374	4.13	8	NZ_AXBB00000000.1
Phenylobacterium zucineum	HLK1	4.38	2	NC_011143.1
Pseudophaeobacter arcticus	DSM23566	5.05	8	NZ_AXBF00000000.1
Rhodobacter capsulatus	SB1003	3.87	2	NC_014035.1
Rhodobacter haeroides	2-4-1	4.60	7	NC_009007.1
Rhodobacter haeroides	ATCC17025	4.56	6	NC_009430.1
Rhodobacter haeroides	ATCC17029	4.49	3	NC_009049.1
Rhodobacter haeroides	KD131	4.71	4	NC_011960.1
Rhodobacterales bacterium	HTCC2083	4.02	5	NZ_DS995280.1
Rhodobacterales bacterium	HTCC2150	3.58	25	NZ_AAXZ01000012.1
Rhodobacterales bacterium	HTCC2255	2.30	2	NZ_DS022282.1

Roseobacter	AzwK3b	4.18	31	NC_008388.1
Roseobacter	CCS2	3.50	11	NC_015730.1
Roseobacter	GAI101	4.53	9	NZ_ABCR01000029.1
Roseobacter	MED193	4.67	4	NZ_AAYB01000010.1
Roseobacter	SK209-2-6	4.56	29	NZ_DS999219.1
Roseobacter denitrificans	OCh114	4.33	5	NZ_AANB01000005.1
Roseobacter litoralis	Och149	4.75	4	NZ_AAYC01000028.1
Roseovarius	217	4.77	6	NZ_CH724156.1
Roseovarius	TM1035	4.21	15	NZ_CH902585.1
Roseovarius nubinhibens	ISM	3.68	4	NZ_ABCL01000004.1
Ruegeria	ANG-R	4.68	41	NZ_JWLJ00000000.1
Ruegeria	ANG-S4	4.54	20	NZ_JWLK00000000.1
Ruegeria	KLH11	4.49	6	NZ_DS999534.1
Ruegeria	R11	3.82	2	NZ_DS999534.1
Ruegeria	TrichCH4B	4.67	129	NZ_DS999055.1
Ruegeria	TW15	4.49	28	NZ_AEYW01000007.1
Ruegeria	TM1040	4.15	3	NC_006569.1
Ruegeria atlantica	CECT4293	4.82	67	NZ_CYPS00000000.1
Ruegeria conchae	TW15	4.49	28	NZ_AEYW00000000.1
Ruegeria halocynthiae	MOLA	4.31	19	NZ_JQEZ00000000.1
Ruegeria lacuscaerulensis	ITI-1157	3.52	47	NZ_FQYJ00000000.1
Ruegeria marina	CGMCC	5.00	53	FMZV00000000.1
Ruegeria mobilis	F1926	4.83	5	NZ_CP015230.1
Rugeria pomeroyi	DSS3	4.60	2	NC_008043.1
Sagittula stellata	E37	5.26	39	NZ_AAYA01000035.1
Sedimentitalea nanhaiensis	DSM24252	4.95	30	NZ_AXBG00000000
Sulfitobacter	EE36	3.37	4	NZ_CH959310.1
Sulfitobacter	NAS-14-1	4.01	11	NZ_CH959313.1
Tateyamaria	ANG-M1	4.43	32	NZ_JWLL00000000
Thalassiobium	R2A62	3.49	1	NZ_GG697169.2
Wenxinia marina	DSM24838	4.18	41	NZ_KB902299.1

Species Name	**Strain**	**Genome Length (Mbp)**	**No. of Scaffolds**	**Asscesion or Number**
Frankia Set				
Cryptosporangium	arvum_44712	9.20	1	NZ_JFBT00000000.1
Cryptosporangium	aurantiacum_DSM_46144	9.58	44	NZ_FRCS00000000.1
Frankia	CcI6	5.58	136	GCA_000503735.2
Frankia	45899	9.54	83	GCA_001536285.1
Frankia	ACN1ag	7.52	90	GCA_001414035.1
Frankia	Allo2	5.35	110	GCA_000733325.1
Frankia	AvcI1	7.74	77	GCA_001420875.1
Frankia	BMG5_23	5.27	166	GCA_000685765.2

Frankia	BMG5_30	5.82	95	GCA_001983005.1
Frankia	BMG5_36	11.20	280	GCA_001854805.1
Frankia	BR_AAY23_1001	5.23	180	GCA_001636575.1
Frankia	Cc1_17	8.36	195	GCA_001854655.1
Frankia	CcI156	5.33	145	GCA_001983015.1
Frankia	CcI49	9.76	78	GCA_001983215.1
Frankia	CED	5.00	120	GCA_000732115.1
Frankia	CgIM4	5.20	135	GCA_001756285.1
Frankia	CgIS1	8.03	289	GCA_001854725.1
Frankia	CN3_FCB_1	9.98	2	GCA_000235425.3
Frankia	coriariae_BMG5_1	5.80	116	NZ_JWIO00000000.1
Frankia	CpI1_P_FF86_1001	7.62	143	GCA_001421075.1
Frankia	CpI1_S_FF36	7.62	153	GCA_000948395.1
Frankia	DC12	6.88	1	GCA_000966285.1
Frankia	Dg2	5.90	2738	GCA_900067225.1
Frankia	discariae_BCU110501	7.89	194	NZ_ARDT00000000.1
Frankia	EAN1pec	8.98	1	NC_009921.1
Frankia	EI5c_UG55_1001	6.62	159	GCA_001636565.1
Frankia	elaeagni_BMG5_12	7.59	135	NZ_ARFH00000000.1
Frankia	inefficax_EuI1c	8.82	1	NC_014666.1
Frankia	EUN1f_ctg00163	9.35	396	GCA_000177675.1
Frankia	EUN1h	9.91	129	GCA_001854645.1
Frankia	Iso899	5.10	67	GCA_000421445.1
Frankia	NRRL_B_16219	5.26	135	GCA_001854695.1
Frankia	asymbiotica_NRRL_B_16386	9.44	174	NZ_MOMC00000000.1
Frankia	QA3	7.59	1	NZ_CM001489.1
Frankia	R43	10.45	46	GCA_001306465.1
Frankia	symbiont_of_Datisca_glomerata	5.34	3	NC_015656.1
Frankia	Thr	5.31	169	GCA_000611815.2
Jatrophihabitans	endophyticus_45627	4.48	10	NZ_FQVU00000000.1
Sporichthya	polymorpha_43042	5.50	1	NZ_AQZX00000000.1
Frankia	alni_ACN14A	7.50	1	NC_008278.1
Frankia	casuarinae_CcI3	5.43	1	NC_007777.1
Frankineae	bacterium_MT45	4.23	1	GCA_900100325.1

TableS03. Genes used in the Frankia MLSA.

Gene	Location on NC_007777.1
glutamate synthase beta subunit	(3573463-3574905)
glutamate synthase alpha subunit	(3574965-3579521)
ribosomal protein large subunit 1	(659953-660669)
large subunit 2	(679418-680251)
large subunit 3	(677637-678344)
large subunit 4	(678341-679078)
ribosomal protein small subunit 1	(1259062-1260540)

small subunit 2	(4280241-4281110)
small subunit 3	(681044-681991)
small subunit 4	(692551-693180)
elongation factor Tu	(675819-677012)
bipA	(4627929-4629767)
ATP synthase subunit alpha	(4442416-4444074)
ATP synthase subunit beta	(4439882-4441321)
dnaA	(35-1723)
dnaK	(5197168-5199018)
dnaX	(309543-311981)
GAPDH	(1969408-1970415)
groEL	(715686-717326)
gyrA	(8155-10659)
gyrB	(6021-7955)
recA	(4210237-4211274)
rpoB	(662999-666424)
rpoD	(1534111-1535289)

Chapter 4 – Rare genes and horizontal gene transfer in the Haloarchaea

This chapter consists of one publication and my progress towards a second. The first publication (Fullmer et al., 2014a) is a book chapter reviewing Horizontal Gene Transfer (HGT) in the Halobacteria. The major themes of the article are that the Halobacteria have a high effective rate of gene transfer, mediated by the traditional methods as well as their possibly unique method of cell-fusion, and that this rate of transfer has shaped their evolution. It was written in collaboration with J. Peter Gogarten and R. Thane Papke. I researched and wrote the manuscript and participated in the editing. J. Peter Gogarten participated in editing and writing the manuscript. R. Thane Papke provided the initial concept and recommended a comprehensive reading list to start the process. Thane also provided much of the direction and supervision to my efforts and participated in the editing and writing of the manuscript.

The 2nd part of this chapter is my progress towards surveying and analyzing restriction-methylation genes in the genus *Halorubrum* and the class Halobacteria as a whole. This work was done in collaboration with R. Thane Papke and J. Peter Gogarten, and in partial collaboration with Matthew Ouellette. Matthew Ouellette assisted in developing the concept and direction of the methylation studies and provided valuable knowledge of the restriction-methylation system. R. Thane Papke and J. Peter Gogarten participated in conception and design of analyses. I participated in the conception and design of analyses and performed all analyses. The major results of this work have been

the identification of 48 candidate restriction-methylation genes and the quantification of their horizontal transfer within the Halobacteria class.

Horizontal Gene Transfer in Halobacteria

3

Matthew S. Fullmer, J. Peter Gogarten and R. Thane Papke

Abstract

The Halobacteria are a class of Archaea that have been fundamentally shaped by Horizontal Gene Transfer (HGT). The mechanisms for HGT are not well understood, or are unreported. A noteworthy exception exists for the genus *Haloferax*, where a novel mating system exists that includes the fusion of cytoplasm between two cells. Despite shallow insight into mechanisms evidence from phylogenetics and population genetics studies demonstrate that these organisms have been able to exchange genes since their distant origins and continue to actively do so today. Single gene studies have uncovered transfer of Halobacterial rhodopsins into diverse lineages such as the fungi and multiple bacterial taxa, construction of novel biosynthetic pathways, homologous recombination of parts or whole ribosomal proteins and RNAs, as well as divergent tRNA synthetases being exchanged between distant lineages. Furthermore, the very origin of the Halobacteria appears to have resulted from an influx of genes from the bacterial domain, which reshaped the fundamental metabolism from an anaerobic chemoautolithotrophic methanogen into a facultative aerobic heterotroph. Population genetics analysis demonstrated that gene flow with phylogenetically defined populations is so frequent that allele distributions resemble that of sexually reproducing eukaryotes, and acts as both a homogenizing and diversifying evolutionary force. Given all of the evidence for abundant recombination into, out of and between these lineages, how then do new, distinct, lineages such as these stably emerge? The answer appears to lie in a balance between recombination as a cohesive force holding populations together as entities recognizable as taxonomic units, and barriers to that transfer for promoting diversification. A primary candidate appears to be geographic barriers that reduce gene transfer between populations sufficiently to allow regional signatures to emerge.

Introduction

The Halobacteria (often referred to colloquially as the Haloarchaea) are a highly recombinant class of salt-loving organisms in the archaeal phylum Euryarchaeota. These remarkable organisms thrive in brines around the globe that have salt concentrations as high as 35% (sea water is ~3.5%). Most strains cannot tolerate concentrations below 10%. Their metabolism is primarily aerobic heterotroph and many garner extra energy by harvesting light via rhodopsin pigments in their membranes. Perhaps most profound, the Halobacteria have been directly constructed or inextricably molded by the exchange of genes between organisms in a non-Mendelian manner.

Mechanisms of horizontal gene transfer

The classic description of HGT includes three mechanisms for transfer of genetic material (Thomas and Nielsen, 2005). These mechanisms are conjugation, transduction and natural transformation. In recent years a fourth process known as gene transfer agents (GTAs) has been elucidated (Lang *et al.*, 2012). Conjugation and transduction are considered mechanisms for the propagation of selfish DNA elements (plasmids and viruses/phage respectively) that only coincidentally transfer their infected host's genes to another cell, while natural transformation most likely evolved to import environmental DNA specifically to benefit the individual cell: it requires many genes and is typically a highly regulated process (Redfield, 2001). GTAs though seemingly more like transduction because DNA is packaged in a virus-like protein coat it is also similar to transformation because it is also a tightly regulated process that specifically moves host DNA from cell to cell. In each of these cases DNA is transferred unidirectionally from a donor to a recipient.

The protein machinery for conjugation is typically encoded by *tra* genes located on conjugative plasmids. These gene products form a specialized pilus which links the cell containing the plasmid (i.e. donor: F′) to a cell that does not (i.e. recipient: F^-). The conjugative plasmids are copied by rolling circle replication, which requires one strand of DNA to be nicked and peeled off while the intact strand is used as a template for DNA synthesis. The non-template strand is then transferred through the pilus into the recipient cell. In the recipient cell the ssDNA is re-circularized and DNA polymerase creates a complementary strand to result in a double-stranded DNA molecule. Sometimes the plasmid will integrate into the host chromosome, or a fraction of the host chromosome will integrate into the plasmid and when the plasmid is transferred it will also bring host genes to the recipient. In these cases, the host chromosome is transferred inadvertently.

Conjugation is a notable unidirectional method of transfer; only cells with the plasmid can act as a donor, and only cells without the plasmid can be recipients. Cells already containing a plasmid cannot receive an additional copy from another donor. Transfer of plasmids can be picky, and promiscuous: sometimes they cannot be transferred to potential donor cells from the same species, yet can infect multiple species.

Transduction is the transfer of genetic material from one cell to another via a viral intermediate. Viruses can pursue two strategies once they have penetrated a host's external defences. The simplest strategy is to enter the lytic cycle. In this process the virus co-opts the host's nucleic acid and protein synthesis machinery for its own purposes while suppressing the host's usage. The virus creates many copies of its genome while also creating new proteinaceous shells (capsids) to transport these genomes to a new host. Lysogenic, or temperate phages as they are also called, do not immediately co-opt the host's molecular machinery for reproduction. Instead, they integrate their genetic material into their host's chromosome. While inserted, they replicate when the host reproduces. Thus, these viruses are able to expand their population through vertical inheritance. Once they detect a signal from the host (usually a stress response) they excise their DNA and begin the lytic phase. When excising their genome and packaging copies into capsids the phage are sometimes imprecise. Some virions are packaged with host DNA. Yet, the virions themselves are fully intact and infectious. Thus, the virions can still transfer their DNA into a new host. Once in the new host the DNA may be recombined into the chromosome through homologous replacement or as part of a non-functional genomic island (the phage 'genome'). The net result is transfer of genetic material from one organism to another in a non-Mendelian fashion.

Transformation is the uptake of DNA from the environment and its integration into the chromosome by a cell (Chen *et al.*, 2005). Natural transformation is intrinsically linked to the process of competence, which is a state that enables cells to transport DNA from the environment across their membrane and into the cell as a high molecular weight molecule. Some bacteria are competent and may use DNA for nutritional purposes but not recombine it into their chromosomes (i.e. not undergo transformation; (Finkel and Kolter, 2001; Redfield, 2001). Movement of DNA from the environment into the cell is complex and requires a set of numerous proteins acting in concert. Many proteins are homologous and conserved between Gram-positive and -negative bacteria yet some of the molecular machinery is more specialized to accommodate the specificities of the cell wall and membrane characteristics. Only few instances of natural competence or transformation have been described for Archaea (Bertani and Baresi, 1987; Johnsborg *et al.*, 2007; Lipscomb *et al.*, 2011; Sato *et al.*, 2003; Worrell *et al.*, 1988).

Gram-negative and -positive model organisms frequently employ mechanisms for biasing the DNA that they import. Pathogenic Gram-negative organisms such as *Neisseria gonorrhoeae* and *Haemophilus influenzae* (among many others) use DNA uptake sequences (DUS) (Smith *et al.*, 1995), which are ~10mer DNA repeats present in high copy number (>1000×) and on both strands of the chromosome, to delineate DNA for import: cell machinery must first recognize a DUS before the DNA can be taken in by the cell. DUSs are typically thought to be species specific, though close relatives can sometimes recognize the same signal. Organisms that utilize DUSs typically constitutively produce competence proteins and can import DNA at any time (Hamilton and Dillard, 2006). This is in contrast to model Gram-positive organisms like *Bacillus subtilis* and *Streptococcus pneumoniae* (Havarstein and Morrison, 1999), which can import any kind of DNA but regulate when they take it up through quorum sensing. This population size-associated timing of protein expression and thus when DNA is imported ensures a large number of their own species are available to donate DNA. Further, the biofilms in which these organisms typically live, diffusional processes like quorum sensing up-regulate competence in only a fraction of the population (Steinmoen *et al.*, 2002). This is relevant because in some instance fratricide is an important aspect of competence: in addition to up-regulating competence, the quorum sensing regulation simultaneously signals for the production of a toxin/antitoxin system (Claverys and Havarstein, 2007). Competent cells produce the toxin and antitoxin while non-competent cells do not. Therefore non-competent cells are lysed and spill out their DNA for the competent cells to utilize.

HGT via gene transfer agents can be broadly thought of as similar to transduction. More specifically, it can be thought of as decayed prophages incapable of packaging their own DNA: the GTA randomly packages host DNA for transfer of DNA to a new host. Furthermore, production of GTAs has been observed to be under the control of a quorum sensing process. While GTA similarity to phage is clear, it is also helpful to think of the process as a 'protected' natural transformation.

Mechanisms of horizontal gene transfer in Halobacteria

The single largest unifying feature of the above mechanisms is the polarity of their transfer. Each of them has a donor from where the genetic material leaves and a recipient to where the material arrives. Thus, a single HGT event can only alter one of the two parties. In contrast,

Haloferax, a genus from the Class Halobacteria (Phylum Euryarchaeota) utilizes a novel mechanism of HGT in which a physical connection is required for transfer, but the donor–recipient relationship is non-existent: all cells are donors and recipients. The evidence for this non-canonical transfer mechanism called mating is multifaceted:

- Shaking cultures rarely if ever produce recombinants. Recombinants are formed readily, however, after cells are pelleted from centrifugation and when cells are incubated on a substratum, indicating the need for cell–cell contact (Mevarech and Werczberger, 1985).
- Nuclease treatment of cells still results in a high frequency of recombinants, indicating natural competence is not involved (Mevarech and Werczberger, 1985).
- Medium filtrate (0.22 μm) from one auxotroph when applied to another does not form recombinants, indicating the process is not virus nor GTA mediated (Mevarech and Werczberger, 1985).
- Auxotrophs form prototrophs in every strain combination ever tried (Mevarech and Werczberger, 1985), whereas in standard conjugation experiments transfer only occurs between specific cell types, thus indicating no donor–recipient relationship for *Haloferax*.
- Plasmid and chromosome-based traits are transferred at the same rate (Naor *et al.*, 2012). In conjugation, chromosomes are much less frequently transferred.
- Electron microscopy reveals the presence of multiple intercellular 'bridges' between cells (Rosenshine *et al.*, 1989), and multiple cells are connected like beads on strings (Mullakhanbhai and Larsen, 1975).
- Cell walls and membranes appear contiguous between cells (Mullakhanbhai and Larsen, 1975). In contrast, conjugation requires the formation of a protein pilus.
- A low concentration of Mg^{2+}, which interferes with the stability of cell membranes, also prevents the formation of bridges (Rosenshine *et al.*, 1989).
- Unusually large segments of chromosomal DNA (10–17% of the total chromosome) are transferred (Naor *et al.*, 2012).
- Plasmids from cells containing multiple plasmids (e.g. *Haloferax volcanii* strain DS2) are transferred in an apparently random process (Naor *et al.*, 2012).
- Mating is a pilin-independent process (Tripepi *et al.*, 2010).

A four-step model has been hypothesized to explain the data (Ortenberg *et al.*, 1999). In step 1, 'haploid' cells grow bridges (pseudopodia-like appendages are seen extending from cells) until they fuse with nearby cells. During step 2, bridges widen to accommodate the flow of cytosolic contents, including chromosomes and plasmids producing 'diploid' or heterochromosomal cells. Step 3 is when recombination of chromosomes occurs, and step 4 is the segregation of chromosomes and plasmids into separate cells. [Quotations around haploid and diploid are used to indicate the inaccuracy of the terms. *Haloferax* spp. and other Halobacteria have recently been determined to be polyploids, containing tens of chromosomes even during stationary phase (Breuert *et al.*, 2006).] This process is remarkably analogous to the haploid/diploid life cycle of sexually reproducing eukaryotes, with the caveat that no reproduction is occurring during the mating activity of Halobacteria. As intriguing as this process is, how this process works and its uniqueness in the biological world remain open questions.

Very little is known about transduction and natural competence in Halobacteria. For instance, viruses are known to infect Halobacterial hosts (e.g. Atanasova *et al.*, 2012; Dyall-Smith *et al.*, 2003), there is evidence for viruses embedded in Halobacterial chromosomes (e.g. Cuadros-Orellana *et al.*, 2007) and virus DNA from hypersaline environments has been sequenced (Rodriguez-Brito *et al.*, 2010), but no virus mediated genetic system has been developed and transduction as a process has not been reported (Allers and Mevarech, 2005) (see also Chapter 4). The genetic systems developed for Halobacteria are based on the formation of spheroplasts for chemical-induced competence (e.g. Charlebois *et al.*, 1987; Cline and Doolittle, 1987), rather than natural competence. Evidence for natural transformation has not been reported.

Evidence for horizontal gene transfer in Halobacteria from single gene studies

Rhodopsins

Even though we do not have a deep understanding for the mechanisms of horizontal gene transfer in Halobacteria, we know it occurs because their chromosomes record ample evidence for rampant HGT. As early as the first sequenced Halobacterial genome, *Halobacterium* sp. strain NRC-1, the respiratory apparatus was identified as having originated in Bacteria (Ng *et al.*, 2000). Subsequent work has reaffirmed the bacterial origin for this system (Nelson-Sathi *et al.*, 2012). The movement of respiration genes into a methanogen-like Halobacterial ancestor probably played a crucial role in the origins of the Halobacteria, and will be discussed in more depth in a following section.

A signature trait of many Halobacteria is the presence of light-harvesting rhodopsins that have three major functions: bacteriorhodopsins (bR), which pump protons out of the cell thus creating an ATP generating proton motive force (Lozier *et al.*, 1975); halorhodopsin (hR), an anion transporter (primarily chloride) for maintaining iso-osmotic balance (Kolbe *et al.*, 2000); and sensory rhodopsin (I and II) to regulate phototaxis: cells move towards useful wavelengths utilized by bR and hR and away from harmful ultraviolet range wavelengths (Spudich and Bogomolni, 1988). Both the Chloride and Sensory paralogues have been reported in HGT events across the domain level. *Salinibacter ruber* contains four rhodopsin homologues. Two of these are sensory in nature and the third is a chloride transporter. Phylogenetic analysis places the *Salinibacter* genes squarely within the Halobacterial clade, suggesting they were recent acquisitions, across domain boundaries, by the bacteria, and not the other way round. Interestingly, a fourth rhodopsin, which is a H^+ pump of no apparent relation to the archaeal version, appears to have been acquired from a divergent bacterial phylum in another HGT event (Sharma *et al.*, 2006).

Fungal rhodopsins also share ancestry with halorhodopsins. Once rapidly evolving sequences that have changed function are screened out of the fungal rhodopsins, it is clear they group with the hR chloride transporters (Sharma *et al.*, 2006). In contrast with the *Salinibacter* example, this appears to be a single transfer event into the ancestor of Fungi as fungal sequences across the entire kingdom group together inside the Halobacteria (Sharma *et al.*, 2006).

Rubrobacter xylanophilus is a member of the Actinobacteria. It is noted for its radiation resistance, and the presence of a rhodopsin (Carreto *et al.*, 1996). Another actinobacterium,

Kinecoccus radiotolerans, has a rhodopsin that is bacterial in origin (Sharma *et al.*, 2006). However, the rhodopsin from *R. xylanophilus* groups within the Halobacterial clade, albeit not clearly within any of the major functional clusters. Both of those rhodopsin proteins fit the primary sequence and secondary structure profiles of proton transporters. Since *R. xylanophilus* is a member of the most basal lineage of the Actinobacteria it was inferred that this transfer occurred after its divergence from the rest of the phylum.

Unsurprisingly, the HGT of rhodopsins is not solely across vast phylogenetic distances. Within the Halobacteria HGT appears to have been a regular occurrence. Rhodopsin genes can be found across the Halobacterial class. However, their distribution is patchy. This has led to suggestions of all four functionalities having existed in the most-recent Halobacterial common ancestor (Ihara *et al.*, 1999). An examination of the specific rhodopsin genes against a species tree finds phylogenetic disagreement (Sharma *et al.*, 2007). Using the rpoB' protein as a marker for vertical descent both the bacteriorhodopsin and halorhodopsin gene trees displayed discordance. Overall, the bacteriorhodopsin gene tree agreed with the 16S rRNA sequence-based species tree but for two instances. These two cases each saw the rhodopsin linked to two adjacent genes, evidentially transferred as a unit. It is not clear if these were two separate events or a single ancestral transfer event. The halorhodopsins appear to have at least two HGT events causing their gene tree to diverge from their bacteriorhodopsin homologues (Sharma *et al.*, 2007).

The linkage of the bacteriorhodopsins with two adjacent genes is worthy of extra note. Sharma *et al.* (2007) found these genes (*bac* and *bap*) to be more frequently linked to bacteriorhodopsins than other genes previously shown to be important in the proton pump function. It is also interesting that in the absence of a bacteriorhodopsin these linked genes are also absent. When this is combined with other observations such as that the *Haloferax volcanii* strain DS2 does not harbour rhodopsins, (it was cultivated from Dead Sea mud, which is not transparent to sunlight, and it utilizes nitrate reduction in conditions unfavourable to rhodopsin use) and the diversification of function in the rhodopsins, one can view rhodopsins as niche-specific functional units that can be lost and gained as the ecology demands (Sharma *et al.*, 2007).

New biochemical pathways assembled through horizontal gene transfer

Halobacteria have built via HGT a new pathway tailored to their environments. In salt lakes, for example, blooms of microorganisms are often rare and short-lived. Halobacteria store carbon in the form of polyhydroxyalkanoate. As a result, pathways for assimilation of carbon may be advantageous. *Haloarcula marismortui* has assembled a novel acetate-assimilating pathway from glutamate fermentation, acetate fermentation and propionate assimilation genes all of which originated from the Bacteria (Khomyakova *et al.*, 2011). The resulting pathway, known as the methylaspartate pathway, allows these halophiles to assimilate acetyl-CoA in microaerobic environments where the canonical glyoxylate pathway cannot function (Khomyakova *et al.*, 2011). The discoverers suggest that this pathway could serve as a mechanism to cope with glutamate overloads.

Ribosomal RNA and proteins

16S rRNA sequences are typically considered the 'gold standard' for phylogenetics, primarily for historical reasons. Early prokaryotic phylogenies that described relationships between

taxa used 16S rRNA gene sequence data (Woese and Fox, 1977; Woese *et al.*, 1990) and the marker became the standard for describing a newly discovered strain's place in the tree of life (Oren *et al.*, 1997; Stackebrandt and Goebel, 1994). With the advent of PCR, environmental 16S rRNA gene sequencing for understanding community composition independently of cultivation vastly expanded the database. After the discoveries of frequent HGT, it was argued that the 16S rRNA gene, and other informational genes, such as those involved in DNA transcription and translation, are resistant to horizontal transfer (Complexity Hypothesis: (Jain *et al.*, 1999)). Transferred genes whose products participated in many macromolecular interactions would be poorly optimized to interact efficiently as part of a complex in their new host. This contrasts with so-called accessory or operational genes that often are involved in few interactions (Jain *et al.*, 1999) and which are frequently gained and lost from lineages (Lawrence and Roth, 1996). For example, a cytosolic catalytic enzyme might not need to interact with any other gene-product to fold, become active and metabolize its substrate. The result is that many accessory genes may be able to function at normal efficiency in foreign environments whereas an informational gene might cause a substantial penalty, and decrease the fitness of the organism it arrived in.

Though the complexity hypothesis is probably enforced most of the time, the examples of apparent ribosomal RNA and ribosomal-associated protein gene transfers have been growing (e.g. Badger *et al.*, 2005; Boucher *et al.*, 2004a; Gogarten *et al.*, 2002; Gupta *et al.*, 2003; Wellner *et al.*, 2007; Williams *et al.*, 2012; Zhaxybayeva *et al.*, 2006, 2009) indicating it is not always deleterious, and perhaps sometimes useful to have a foreign-derived informational gene. Specific examples can be found in the Halobacteria. *H. marismortui* possesses two rRNA operons. The SSU genes are 5.0% divergent from each other while the large subunit (LSU) genes are 1.3% dissimilar (Mylvaganam and Dennis, 1992). *Halosimplex carlsbadense* possesses SSU genes 6.8% divergent and LSUs 2.7% divergent; *Natrinema* sp. strain XA3-1 exhibits four operons wherein three SSU genes are ~0.1% distant and the fourth is ~5.0%; all four LSUs are ~1–2% distant; in both *Har. marismortui* and *Hsx. carlsbadense* the rRNA intergenic spacers (ITS), are 24.6% and 49.1% divergent, respectively; the *Natrinema* strain displays zero divergence in its ITS region (Boucher *et al.*, 2004a).

Each *Har. marismortui*, *Hsx. carlsbadense* and *Natrinema* SSU gene clearly does not share a common recent history with its intragenomic relatives. In *Natrinema* the three nearly identical copies fit expectations but the fourth with ~5.0% divergence did not display strong affinity with any other known species at the time. In the other two strains, the very high level of divergence between the two rRNA operons suggest that at least one version likely evolved in a different lineage and was horizontally transferred into its current genome.

The *Natrinema* sequences are particularly interesting. The ITS regions are among the most hyper-variable and should diverge much faster than the rRNA genes themselves, which have large amounts of their sequence under strong purifying selection. However, all four are identical. It is possible that the ITS locus is under strong purifying selection though no function is known; more likely recombination between the rRNA operons coexisting in the same genome recently purged the diversity in this region, as was discussed for the rRNA operons in *Thermomonospora chromagena* (Gogarten *et al.*, 2002; Yap *et al.*, 1999). In contrast, the LSU genes themselves have all accrued ~1.0% separation from each other.

A more detailed examination of the evolutionary signals within the divergent genes identified an apparent transfer between *Natrinema* sp. XA3–1 strain and *Natrialba magadii*. A 100-bp region from one *Natrinema* copy showed strong similarity to the corresponding

region in *N. magadii*'s LSU genes (Boucher *et al.*, 2004a). This indicates that the region was transferred into *Natrinema*, a homologous recombination event occurring across a distance of ~5%, while the remainder of the LSU was not.

There are many problems for microbial phylogeneticists, ecologists and taxonomists associated with highly divergent intragenomic rRNA heterogeneity. First, it is greater than that typically seen within species, which usually is below ~2.0%. This creates difficulties for the suitability of rRNA genes as a marker for classification in strains featuring such heterogeneity: which SSU copy is to be used for determining relationships? Further, with divergent copies in a single chromosome, the difficulties of obtaining a reliable sequence are immense. Multiple divergent copies will cause a PCR product that is directly sequenced (i.e. without cloning first) to have dozens of sites across the alignment be unresolved, with peaks in sequencing chromatograms for two different nucleotides at the same position. If the PCR product is cloned and then sequenced, chimera artefacts will likely occur leading to overestimating the number of copies, and sequence variation per copy. For example, when *Halosimplex carlsbadense* was first described, it reportedly contained three distinct 16S rRNA genes, with A and B being 97.7% similar and C being 93.8 and 92.2% similar to A and B respectively (Vreeland *et al.*, 2002). Later, using PCR independent techniques, it was demonstrated that *Hsx. carlsbadense* has only two distinct copies that are 6.8% divergent (Boucher *et al.*, 2004a). Multiple divergent operons from single cells will also over estimate community species richness and abundance (in addition to forming PCR induced chimeric sequences) when applying standard PCR and cloning techniques to analysis of environmental DNA from saturated brines. Finally, fractions of rRNA genes can be transferred independently of the entire gene or operon (Boucher *et al.*, 2004a). Thus, a single recombination event might sufficiently obscure the 'true' evolutionary signal. A single transfer of several hundred nucleotides might be spotted, if multiple sequence alignments are analysed; however, mosaics formed as the result of multiple gene conversion events that ultimately lead to homogenization of the initially divergent copies coexisting in a genome likely remain undetectable in most instances.

Gene conversion, the effect of homologous recombination between two copies of the same gene on the same chromosome, is known to occur in Halobacteria (Lange *et al.*, 2011) and would prevent divergence of loci (Liao, 1999) unless subfunctionalization occurs and variation is selected for. Without subfunctionalization and selection, there would likely be no bias towards maintaining the original copy and both copies are expected to contribute to the homogenized mosaic resulting from many gene conversion events. In many Halobacteria, rRNA heterogeneity appears stably maintained and may be quite common within the class (Dennis *et al.*, 1998; Grant *et al.*, 1998). Evidence from analysis of rRNA operons in *Haloarcula* and *Halomicrobium* species has demonstrated that the divergent copies are expressed under different laboratory conditions, e.g. temperature and salinity, and probably confer adaptive advantages when the organism finds itself in fluctuating environmental circumstances (Cui *et al.*, 2009; Lopez-Lopez *et al.*, 2007).

tRNA synthases

The Halobacteria have some tRNA synthases with unusual evolutionary histories. The class contains two versions of the Leucine tRNA synthase (Andam *et al.*, 2012). The first type is the archaeal version that falls within the Euryarchaeota (LeuRS-A). The other type, more common among the Halobacteria, is more similar to bacterial homologues (LeuRS-B). In

phylogenetic reconstruction the B-type does not actually group inside the cluster of bacterial homologues. Rather, it branches deeper than the last common bacterial ancestor. Thus, it appears that the Halobacteria received the LeuRS-B genes from an unsampled, or possibly extinct, lineage that diverged from the Bacteria before that domain's common ancestor appeared.

The Halobacterial LeuRS gene distribution is the product of HGT. When the presence of LeuRS-A and B are mapped onto an MLSA tree, created using concatenated housekeeping genes, an unusual pattern emerges. All of the strains that possess an A-version lack a B-version. Likewise, all of the strains with a B-version do not have an A-version. What is striking is that the A and B gene type phylogenetic trees and the MLSA derived versions are conflicted, and carriers do not form monophyletic groups. Apparently, several lineages have replaced their own LeuRS genes with extremely divergent out-paralogues (xenologues). The LeuRS A and B loci are not in syntenic gene neighbourhoods suggesting that lineages gained the xenologous copy through non-homologous recombination and that after co-existing for some time one or the other LeuRS version was lost.

The possibility that lineages maintained the two divergent LeuRSes is possibly supported by the evolutionary dynamics of the B-version. The B-version has a split inside its own group, which divides it into two subgroups, designated B′ and B″. Every strain with a B″ copy is also a carrier of B′. However, B′ is often a solitary copy. There are two possible explanations for this arrangement. The first is that in taxa with both versions the protein has evolved to function as a hetero-dimer. Second, this could be evidence that multiple versions can be maintained for evolutionarily meaningful time before differential loss purges one or the other. Presumably, if this is the case, both copies confer some form of selective advantage under different conditions to enforce their maintenance (Andam *et al.*, 2012).

The origins of the Halobacteria may be rooted in horizontal gene transfer

The different studies above describing the number of gene families that have been transferred into or among the Halobacteria suggests that this class of organisms have been heavily impacted by participation in HGT events throughout their evolutionary history. New evidence is now suggesting that the very origins of the Halobacteria are founded on the influx of genes from elsewhere, especially from the Bacteria.

Phylogenetic analysis, including that of the 16S rRNA gene and ribosomal proteins, place the Halobacteria as sister to *Methanosarcina*, and unrelated to *Methanococcus* and *Methanobacterium* (Matte-Tailliez *et al.*, 2002). In contrast, phylogenies drawn completely from the presence and absence of genes across the entire Archaea place the Halobacteria near the root of the domain (Korbel *et al.*, 2002). The methanogen placement has gained overall acceptance. The supposition that the ancestor to all Halobacteria was a methanogen is rather interesting, as they are dissimilar in almost every metabolic pathway. Methanogens are obligate anaerobes and chemoautolithotrophs. They create methane from H_2 and CO_2 to fuel synthesis of ATP. They also use the acetyl-CoA pathway to convert CO_2 into cellular material. Halobacteria are obligate heterotrophs. They are also either obligate or facultative aerobes. Many supplement energy production by generating a proton gradient via their light driven proton pump (bacteriorhodopsins). However, this is not an essential component to their growth and live exclusively from oxidizing organic carbon substrates (Oren, 2008).

Both the Halobacteria and their Methanosarcinales neighbours show large bacterial signatures in their genomes (Nelson-Sathi *et al.*, 2012). In the case of the Halobacteria, as many as 1089 genes (of 1479 common to at least two Halobacterial genomes from a dataset of 10) place the Halobacterial version as branching within or next to the Bacteria. Among these genes are those required to transform Halobacteria into the oxidative heterotrophs they are today. Critical are the genes encoding the electron transport chain. In total, almost half (482, 44%) of the putative transfers from bacteria are related to metabolism (Nelson-Sathi *et al.*, 2012). Several genes in the isoprenoid biosynthesis pathway are also among these. IDI1, a type 1 isopentenyl diphosphate isomerase is an analogue of the ubiquitous Archaeal version but was acquired from the Bacteria and the Halobacterial glycerol dehydrogenase proteins group within the bacterial clade (Boucher *et al.*, 2004b).

The question then arises whether or not these genes arrived largely simultaneously or as piecemeal acquisitions. Genes present in all ten genomes (473 total) displayed similar topologies within their phylogenetic trees that match the topology of a tree generated from 56 universally distributed archaeal genes, as confirmed by a goodness of fit test. This agreement of topology suggests that these genes share a common vertically descended history. Thus, a considerable fraction of the bacterially acquired genes appear to have arrived at the root of the class, prior to the genetic radiation of the modern Halobacteria (Nelson-Sathi *et al.*, 2012). Nelson-Sathi *et al.* (2012) then boldly suggest that the transferred genes came from a single bacterial donor, similar in concept to the acquisition of mitochondria and thus the origin of eukaryotes popularized by the Hydrogen Hypothesis (Martin and Muller, 1998). However, in the case of the Halobacteria, transfer was ultimately a failed attempt at eukaryogenesis, resulting in 'just' aerobic heterotrophic Archaea. While certainly daring and thought provoking, the hypothesis of a single bacterial donor is not supported by the data.

Homologous recombination within and between Halobacterial lineages

As the studies above demonstrate, recombination is an important force shaping the Halobacteria. This force profoundly affects how their taxonomy and species must be thought of. It has been known for more than 15 years that prokaryotes do not recombine at the same rate with every other lineage of prokaryotes. More specifically, a log-linear distance–decay relationship has been observed between the recombination rate and the sequence divergence between the two organisms (Fraser *et al.*, 2007; Vulic *et al.*, 1997; Williams *et al.*, 2012). Two members of the same species might recombine at a very high rate whereas individuals from different genera would do so at a much lower rate. The result being that transfers across phyla or domains are exceedingly uncommon. The Halobacteria have been demonstrated to have a log-linear distance–decay relationship for homologous recombination, but it appears to be less of a barrier than that seen in bacterial model organisms (Naor *et al.*, 2012; Williams *et al.*, 2012).

HGT has been proposed as a homogenizing force upon populations of related cells (Andam *et al.*, 2010a; Doolittle and Zhaxybayeva, 2009; Feil *et al.*, 2000; Gogarten *et al.*, 2002; Lawrence, 2002; Papke, 2009; Smith *et al.*, 2000; Whitaker *et al.*, 2005). Contrary arguments have been put forth that the level of HGT required to homogenize a population is much higher than that observed in nature (Cohan, 2002, 2006). The question then centres on how much recombination is required to create the phylogenetic clustering or

'clumpiness' observed in nature. Fraser *et al.* (2007) explored some of the impacts of varying recombination rates via a neutral (i.e. selectionless) model. When recombination rates are far below the mutation rate clusters of related organisms form. These clusters are composed of individual cells that can be thought to have their own reproductive fate. They behave in a manner similar to that posited by ecotype-style hypotheses (Cohan, 2006). When recombination is much higher than the mutation rate a different scenario emerges. The diversity of alleles remains the same. However, the number of unique genotypes is higher than in the low recombination situation. The clustering seen under the high recombination to mutation regime is more pronounced but mostly transient. However, when the recombination ratio is between 0.25 and four times that of the mutation rate coherent populations are easily discerned and stable. Two extraordinary conclusions, which are in direct conflict with the ecotype hypothesis, are drawn from this analysis: (1) recombination can act as a cohesive force at biologically observed recombination rates and (2) selection is not required to create the appearance of unique genetic clusters (i.e. species): random birth/death processes are sufficient (Fraser *et al.*, 2007).

The first consequence of recombination occurring at the observed rates is that, as mentioned above, it can act as a cohesive force for populations of organisms (Andam and Gogarten, 2011; Andam *et al.*, 2010b; Gogarten *et al.*, 2002). These clustered populations can be thought of as species, although that definition is still quite inadequate, as will be discussed below. The recombination-dictated clustering, combined with the distance-decay frequencies for recombination, suggest that many gene phylogenies within a 'species' will be conflicted while those outside the species will not be so – although their own intra-clustering trees will be conflicted, too (Dykhuizen and Green, 1991; Lawrence, 2002). The second consequence is a lack of clonality within a species, despite asexual reproduction. As genes are transferred between organisms new combinations of alleles will be created that may have been impossible, or at least statistically impossible, without recombination. This has already been observed via computer modelling (Fraser *et al.*, 2007), and is similar to the 'shuffling of a deck of cards' metaphor used by Dawkins (1976) in his book *The Selfish Gene* to describe sexual recombination. If one were playing a game of cards where each player had a hand and could only change their hand one card at a time, in gradual intervals, from a draw pile it might be thought of as analogous to the situation predicted by an ecotype hypothesis. Here, players would be largely stuck with the hand they had been dealt and when a player is forced to fold the diversity (cards) they possess is lost to the population (card players). Each card in a player's hand would be linked to the fate of the hand as a whole, no matter how advantageous that card may be. However, if the game allowed players to swap cards with each other, either through a communal pool or by directly taking cards from another player's hand, players could radically alter their hands in short order and sample new card combinations which may prove fruitful. Each card would also find itself less linked to the fate of the hand (player) as a whole. Even if a player were to fold, a strong card from their hand may have already found new life in another player's hand.

How do the Halobacteria fare compared to these predictions and consequences developed for bacterial models? Multilocus sequence analysis of over 150 *Halorubrum* strains cultivated from three sites (two different salinity ponds from a saltern in Santa Pola, Spain, and a salt lake in Algeria) demonstrated that recombination was so frequent within phylogenetically defined clusters of closely related strains (< 1% DNA divergence; called phylogroups) that alleles at loci were unlinked to the extent that there was no association of

alleles and loci (i.e. there was no evidence for clonality) (Papke *et al.*, 2004, 2007). Measurements of recombination to mutation demonstrated that loci change twice as frequently by recombination than mutation, which is right in the middle of the range identified by Fraser *et al.* (2007) – see above. The two above observations for random association of alleles and loci, and a rate of homologous recombination in agreement with computer model studies, strongly supports that *Halorubrum* clusters are homogenized by homologous recombination rather than by a selective sweeps model proposed for Cohan's ecotype hypothesis.

Because genes were unlinked through recombination, if selection were applied to a single allele, in theory it could rise in frequency independently of the genes in the rest of the chromosome. Evidence for the fixation of a single allele was seen in the bacteriorhodopsin locus for two *Halorubrum* phylogroups. The *bop* showed minimal nucleotide diversity; approximately 90% of the strains had the same allele in their respective phylogroup, and the only mutations seen were in two strains where A and T mutations were observed in neutral third codon positions. These are likely transient mutations as Halobacteria are well know for their high G + C genomic content, which is especially pronounced at 3rd codon positions, reaching approximately 90%. The other loci examined in that study all demonstrated more polymorphisms in 3rd codon positions, in some cases nearly 10 times more, and for each of the additional loci, no allele was found in more than 50% of the strains. The lack of diversity at the third codon positions, which due to the redundancy in the genetic code can be assumed to be neutrally evolving at least within phylogenetic clusters, while maintaining diversity at other loci in the third codon position is indicative of a recent fixation of the *bop* allele only. If frequent recombination did not occur and genes remained linked, then selection on one adaptive allele that led to its fixation would have also fixed all the other loci in the population, and all diversity within the *Halorubrum* phylogroups would have been purged. Because diversity was purged in only one locus, the only possible explanation is that there is frequent recombination that enables the fixation of single alleles.

Another discussion point to be made from that study is the appearance of well-supported clusters. Clustering is often considered evidence for species, which of course has a lot of intellectual baggage, one of which would be from the Darwinian model that two sister species have a common ancestor. However, that outcome was not necessarily observed. For instance, despite the fact that individuals typically generated the same clusters, irrespective of the locus examined, each gene supported a different sister lineage, indicating the lack of a common ancestor for those 'species'. The only reasonable explanation is that genes originating from outside the population are continuously homologously recombined and fixed in the population, in the same way as described above for the *bop* genes. Furthermore, the 16S rRNA gene tree was completely incongruent with the concatenated gene tree, indicating it is probably the most frequently recombined gene among closely related phylogenetic clusters, an outcome of its extreme sequence conservation and the log-linear frequency of recombination and genetic distance relationship. Interesting to point out is that it is not the strain relatedness but the gene relatedness that is important for frequency of recombination, indicating that numerous recombination events could be ongoing at multiple loci between multiple 'lineages', simultaneously (Papke *et al.*, 2007).

Given that recombination is the diving force behind diversification (and by extension selection) in the Halobacteria, it is logical to attempt to quantify this level of influence. Recent studies have confirmed that the class as a whole is highly recombinogenic (Naor *et al.*, 2012; Williams *et al.*, 2012). Both Williams *et al.* (2012) and Naor *et al.* (2012) recover

a distance–decay relationship between relatedness and frequency of recombination. In Williams' case the frequency of recombination and genetic distance is measured using ribosomal protein tree as a query tree to discover recombined genes on the chromosomes and ribosomal protein divergence to estimate how related the donor and recipient are. Naor measured directly the frequency of recombination using genetically manipulatable strains with sequenced genomes: *Hfx. volcanii*, and *Hfx. mediterranei*. Williams found the log-linear relationship to hold throughout the entire class (using 21 genomes, representing all of the major Halobacterial groups). Examining the relaxed core (defined as genes found in 15 of the 21 Halobacterial genomes) it was estimated that 11–20% of genes evolved in other taxa. Naor's work focused on putative barriers to mating (cell fusion) and recombination within and between species that display approximately 14% nucleotide divergence. As mentioned above, mating is a multistep process that first involves cell membrane fusion, and then recombination. The rate of cell fusion (a 'premating' barrier) was measured by taking advantage of the fact that mated cells remain in a heterochromosomal (diploid) state that preserves the presence of molecular markers from both of the mated strains: PCR amplification was used to assess the presence of each molecular marker. The numbers of between species cell fusion events was less than an order of magnitude smaller compared to within species cell fusions. After measuring the successful fusion events, selection pressure was added to heterochromosomal colonies to identify any hybrid strains that underwent recombination, and an estimate for recombination (a 'postmating' barrier) frequency was made. Surprisingly, the difference within to between species recombination was less than an order of magnitude different. The results recapitulated the expected distance–decay relationship. However, the slope was drastically reduced compared to expectations (e.g. see Vulic *et al.*, 1997; Zawadzki *et al.*, 1995). Instead of the multifold orders of magnitude drop-off in recombination seen in bacterial data (and the basis for computer simulations of Fraser *et al.*), less than an order of magnitude decrease was observed. However, these data still support a 'clumpiness' to organisms in nature defined by groups or clusters that engage in more recombination amongst each other than between groups, something we refer to as preferred trading partners (Papke and Gogarten, 2012).

Geographic isolation and barriers to recombination

From the evidence, it becomes fairly easy to conclude that homologous recombination is an overwhelming homogenizing force for maintaining genetic cohesion within populations of gene trading partners. New alleles that provide a selective advantage arising from mutation can sweep through populations without affecting even nearby loci on a chromosome. In comparison, if loci were linked on chromosomes, the selection process would affect all loci equally bringing them all to zero diversity. Most data indicate that loci have varying amounts of sequence variation (especially at 3rd codon positions), the best explanation for this observation is that genes are unlinked due to recombination. Alleles that originate from outside a recombining population and provide an adaptive advantage will have the same fate as one that arose by mutation alone. However, it will leave a 'foreign' signature because the phylogenetic reconstruction will show a different relationship among relatives for the same gene. When new alleles are invading populations on a regular basis, the evolutionary history of the group will acquire a mosaic appearance, in comparison to any relatives. In a recent publication on marine *Vibrio* populations, comparative genome analysis demonstrated that

a very small fraction (approximately only 100 loci out of thousands shared) had the same phylogenetic topology, indicating that genes were invading and being fixed in populations on a regular basis (Shapiro *et al.*, 2012). Within Halobacterial populations, gene flow is fast enough to unlink alleles and loci to the point of random association, to fix alleles in populations and to generate every possible evolutionary relationship for individual genes on a chromosome (Papke *et al.*, 2007). The important point to emphasize here is that there is an additional layer of complexity between diverged strains: the rate of homologous recombination though slower between than within phylogenetic clusters, can still be quite high, dictating that once diverged, two populations could re-merge given an increase in trading frequency (for interesting possibilities see Sheppard *et al.*, 2008; Zhaxybayeva *et al.*, 2009). The conclusion drawn is that preventing recombination between closely related cells (e.g. sister clones) is very difficult. Yet divergence occurs nonetheless! The question, then, is 'How?' We suggest a role for geographic isolation.

The geographic distribution of the Halobacteria may play a major role in the dynamic interplay between homogenizing and diverging forces. A fundamental fact is that the same genera, and some might argue species within the Halobacteria have a global distribution (e.g. *Haloquadratum walsbyi* (Dyall-Smith *et al.*, 2011; Oh *et al.*, 2010)). Contrasted against this is the reality that not every halophilic environment is the same (e.g. thalassohaline vs. athalassohaline; basic vs. neutral pH). Examination of community composition for solar saltern saturated brines (~35%NaCl) located in Alicante, Spain and Chula Vista, California, revealed dramatically different species richness and abundance, despite both sites being sampled at the same time, being derived from seawater, and existing at the same latitude (Zhaxybayeva *et al.*, 2013). Alicante hosted almost completely Halobacterial inhabitants, and primarily the genus *Haloquadratum*, while Chula Vista possessed an approximately 50:50 split between archaeal and bacterial communities. Each site was uniquely constructed, and had limited shared operational taxonomic units even when applying 95% identity definitions, suggesting enormous difficulty for most halophilic prokaryotes to freely disperse across great distances. Interestingly, saturated brines from Australia that have similar chemical composition also displayed clear difference in species richness and abundance between locations, and very few shared OTUs in comparison to site-specific ones (Oh *et al.*, 2010), indicating even within continents it is difficult to disperse between sites. In the case for strains from the same species, evidence for geographic affect is also found. Comparison of two *Haloquadratum walsbyi* genomes from Spain and Australia demonstrated approximately 1.4% sequence divergence across roughly 84% of their shared genomic regions (Dyall-Smith *et al.*, 2011). Given the known high recombination rates of Halobacteria, which act as a homogenizing force, it is unlikely these two isolates have lived in the same location for some time.

Despite finding a few recognizable Halobacterial OTUs across the globe, the evidence indicates that each halophilic site is unique in composition. This on the surface seems paradoxical, as with the ability to freely disperse, one would expect that each site should be composed similarly. There could be at least two explanations for the observation. Sampling could be a big issue: how deeply sites are sampled could affect the observations and conclusions made. If methods cannot obtain a statistically significant representation of the true diversity and its abundance than we are in deep trouble as ecologists and evolutionists. Molecular techniques though having limitations and biases seem to be very sensitive for estimating diversity and abundance, however. Underlying some of the objection to the observation of differentially composed communities, and the push for the 'poor sampling'

hypothesis is the culturally ingrained and ancient microbiological dictum 'everything is everywhere', which is taken to mean that if two distant locations have similar abiotic conditions they should be composed of the same microbiota (i.e. no dispersal or invasiveness limitations). This idea came out of the Darwinian revolution (i.e. only fitness affects speciation) and before ideas of geographic endemism were accepted for macroflora and fauna (O'Malley, 2007). Add to this layer a dearth of technological knowhow for most of the twentieth century for identifying prokaryotic diversity, and a real possibility that we may never know what bacterial species are, it is no wonder that we can see the 'same thing' everywhere. In the world of macrofauna and flora biologists, geographic isolation is the null hypothesis for speciation, and proving deviations from the null model is required to propose sympatric speciation. Though sympatric speciation occurs in animals and plants, only a few rare conditions exist that can promote it in freely recombining populations (Hey and Pinho, 2010). We microbiologists do not have a null model for speciation, but we seem to accept sympatric speciation frequently without examining the global distribution. Data are accruing in favour of geographic endemism for halophiles and Halobacteria. The same types of observation are frequently made: similar species and genera are found in many places but often there are radiations of microdiversity in specific locations, and that similar sites are composed of different diversity. We think the data support a hypothesis that periods of endemism occur in sites all over the world, but that it happens over short geological time. Short durations of endemism are followed by dispersal events. If recombination prevents divergence in populations, and there is excellent evidence indicating that this is the rule, not the exception, but even if it is not and the homogenizing force is selective sweeps (e.g. see Papke and Ward, 2004), then migration to a new location will counteract both of those forces and divergence can ensue. This means that speciation will be easily facilitated if the rate of mutation and recombination between 'species' is faster as an evolutionary force at generating diversity than the rate of dispersal is at finding places already occupied by the same 'species'.

References

Allers, T., and Mevarech, M. (2005). Archaeal genetics – the third way. Nat. Rev. Genet. *6*, 58–73.

Andam, C.P., and Gogarten, J.P. (2011). Biased gene transfer and its implications for the concept of lineage. Biol. Direct *6*, 47.

Andam, C.P., Harlow, T.J., Papke, R.T., and Gogarten, J.P. (2012). Ancient origin of the divergent forms of leucyl-tRNA synthetases in *Halobacteriales*. BMC Evol. Biol. *12*, 85.

Andam, C.P., Williams, D., and Gogarten, J.P. (2010a). Biased gene transfer mimics patterns created through shared ancestry. Proc. Natl. Acad. Sci. U.S.A. *107*, 10679–10684.

Andam, C.P., Williams, D., and Gogarten, J.P. (2010b). Natural taxonomy in light of horizontal gene transfer. Biol. Philos. *25*, 589–602.

Atanasova, N.S., Roine, E., Oren, A., Bamford, D.H., and Oksanen, H.M. (2012). Global network of specific virus–host interactions in hypersaline environments. Environ. Microbiol. *14*, 426–440.

Badger, J.H., Eisen, J.A., and Ward, N.L. (2005). Genomic analysis of *Hyphomonas neptunium* contradicts 16S rRNA gene-based phylogenetic analysis: implications for the taxonomy of the orders *'Rhodobacterales'* and *Caulobacterales*. Int. J. Syst. Evol. Microbiol. *55*, 1021–1026.

Bertani, G., and Baresi, L. (1987). Genetic transformation in the methanogen *Methanococcus voltae* PS. J. Bacteriol. *169*, 2730–2738.

Boucher, Y., Douady, C.J., Sharma, A.K., Kamekura, M., and Doolittle, W.F. (2004a). Intragenomic heterogeneity and intergenomic recombination among haloarchaeal rRNA genes. J. Bacteriol. *186*, 3980–3990.

Boucher, Y., Kamekura, M., and Doolittle, W.F. (2004b). Origins and evolution of isoprenoid lipid biosynthesis in archaea. Mol. Microbiol. *52*, 515–527.

Breuert, S., Allers, T., Spohn, G., and Soppa, J. (2006). Regulated polyploidy in halophilic archaea. PLoS One *1*, e92.

Carreto, L., Moore, E., Fernanda-Nobre, M., Wait, R., Riley, P., Sharp, R.J., and Da Costa, M.S. (1996). *Rubrobacter xylanophilus* sp. nov., a new thermophilic species isolated from a thermally polluted effluent. IJSEM *46*, 460–465.

Charlebois, R.L., Lam, W.L., Cline, S.W., and Doolittle, W.F. (1987). Characterization of pHV2 from *Halobacterium volcanii* and its use in demonstrating transformation of an archaebacterium. Proc. Natl. Acad. Sci. U.S.A. *84*, 8530–8534.

Chen, I., Christie, P.J., and Dubnau, D. (2005). The ins and outs of DNA transfer in bacteria. Science *310*, 1456–1460.

Claverys, J.P., and Havarstein, L.S. (2007). Cannibalism and fratricide: mechanisms and *raisons d'etre*. Nat. Rev. Microbiol. *5*, 219–229.

Cline, S.W., and Doolittle, W.F. (1987). Efficient transfection of the archaebacterium *Halobacterium halobium*. J. Bacteriol. *169*, 1341–1344.

Cohan, F.M. (2002). Sexual isolation and speciation in bacteria. Genetica *116*, 359–370.

Cohan, F.M. (2006). Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. Philos. Trans. R. Soc. Lond. B Biol. Sci. *361*, 1985–1996.

Cuadros-Orellana, S., Martin-Cuadrado, A.B., Legault, B., D'Auria, G., Zhaxybayeva, O., Papke, R.T., and Rodriguez-Valera, F. (2007). Genomic plasticity in prokaryotes: the case of the square haloarchaeon. ISME J. *1*, 235–245.

Cui, H.L., Zhou, P.J., Oren, A., and Liu, S.J. (2009). Intraspecific polymorphism of 16S rRNA genes in two halophilic archaeal genera, *Haloarcula* and *Halomicrobium*. Extremophiles *13*, 31–37.

Dawkins, R. (1976). The Selfish Gene (Oxford University Press, Oxford, UK).

Dennis, P.P., Ziesche, S., and Mylvaganam, S. (1998). Transcription analysis of two disparate rRNA operons in the halophilic archaeon *Haloarcula marismortui*. J. Bacteriol. *180*, 4804–4813.

Doolittle, W.F., and Zhaxybayeva, O. (2009). On the origin of prokaryotic species. Genome Res. *19*, 744–756.

Dyall-Smith, M., Tang, S.L., and Bath, C. (2003). Haloarchaeal viruses: how diverse are they? Res. Microbiol. *154*, 309–313.

Dyall-Smith, M.L., Pfeiffer, F., Klee, K., Palm, P., Gross, K., Schuster, S.C., Rampp, M., and Oesterhelt, D. (2011). *Haloquadratum walsbyi*: limited diversity in a global pond. PLoS One *6*, e20968.

Dykhuizen, D.E., and Green, L. (1991). Recombination in *Escherichia coli* and the definition of biological species. J. Bacteriol. *173*, 7257–7268.

Feil, E.J., Smith, J.M., Enright, M.C., and Spratt, B.G. (2000). Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. Genetics *154*, 1439–1450.

Finkel, S.E., and Kolter, R. (2001). DNA as a nutrient: novel role for bacterial competence gene homologs. J. Bacteriol. *183*, 6288–6293.

Fraser, C., Hanage, W.P., and Spratt, B.G. (2007). Recombination and the nature of bacterial speciation. Science *315*, 476–480.

Gogarten, J.P., Doolittle, W.F., and Lawrence, J.G. (2002). Prokaryotic evolution in light of gene transfer. Mol. Biol. Evol. *19*, 2226–2238.

Grant, W.D., Gemmell, R.T., and McGenity, T.J. (1998). Halobacteria: the evidence for longevity. Extremophiles *2*, 279–287.

Gupta, R.S., Pereira, M., Chandrasekera, C., and Johari, V. (2003). Molecular signatures in protein sequences that are characteristic of Cyanobacteria and plastid homologues. Int. J. Syst. Evol. Microbiol. *53*, 1833–1842.

Hamilton, H.L., and Dillard, J.P. (2006). Natural transformation of *Neisseria gonorrhoeae*: from DNA donation to homologous recombination. Mol. Microbiol. *59*, 376–385.

Havarstein, L.S., and Morrison, D.A. (1999). Quorum sensing and peptide pheromones in streptococcal competence for genetic transfromation. In Cell–Cell Signaling in Bacteria, Dunny, G.M., and Winans, S.C., eds. (ASM Press, Washington DC).

Hey, J., and Pinho, C. (2010). Divergence with gene flow: models and data. Annu. Rev. Ecol. Evol. Syst. *41*, 215–230.

Ihara, K., Umemura, T., Katagiri, I., Kitajima-Ihara, T., Sugiyama, Y., Kimura, Y., and Mukohata, Y. (1999). Evolution of the archaeal rhodopsins: evolution rate changes by gene duplication and functional differentiation. J. Mol. Biol. *285*, 163–174.

Jain, R., Rivera, M.C., and Lake, J.A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. Proc. Natl. Acad. Sci. U.S.A. *96*, 3801–3806.

Johnsborg, O., Eldholm, V., and Havarstein, L.S. (2007). Natural genetic transformation: prevalence, mechanisms and function. Res. Microbiol. *158*, 767–778.

Khomyakova, M., Bukmez, O., Thomas, L.K., Erb, T.J., and Berg, I.A. (2011). A methylaspartate cycle in haloarchaea. Science *331*, 334–337.

Kolbe, M., Besir, H., Essen, L.O., and Oesterhelt, D. (2000). Structure of the light-driven chloride pump halorhodopsin at 1.8 A resolution. Science *288*, 1390–1396.

Korbel, J.O., Snel, B., Huynen, M.A., and Bork, P. (2002). SHOT: a web server for the construction of genome phylogenies. Trends Genet. *18*, 158–162.

Lang, A.S., Zhaxybayeva, O., and Beatty, J.T. (2012). Gene transfer agents: phage-like elements of genetic exchange. Nat. Rev. Microbiol. *10*, 472–482.

Lange, C., Zerulla, K., Breuert, S., and Soppa, J. (2011). Gene conversion results in the equalization of genome copies in the polyploid haloarchaeon *Haloferax volcanii*. Mol. Microbiol. *80*, 666–677.

Lawrence, J.G. (2002). Gene transfer in bacteria: speciation without species? Theor. Pop. Biol. *61*, 449–460.

Lawrence, J.G., and Roth, J.R. (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics *143*, 1843–1860.

Liao, D. (1999). Concerted evolution: molecular mechanism and biological implications. Am. J. Hum. Genet. *64*, 24–30.

Lipscomb, G.L., Stirrett, K., Schut, G.J., Yang, F., Jenney, F.E. Jr., Scott, R.A., Adams, M.W., and Westpheling, J. (2011). Natural competence in the hyperthermophilic archaeon *Pyrococcus furiosus* facilitates genetic manipulation: construction of markerless deletions of genes encoding the two cytoplasmic hydrogenases. Appl. Environ. Microbiol. *77*, 2232–2238.

Lopez-Lopez, A., Benlloch, S., Bonfa, M., Rodriguez-Valera, F., and Mira, A. (2007). Intragenomic 16S rDNA divergence in *Haloarcula marismortui* is an adaptation to different temperatures. J. Mol. Evol. *65*, 687–696.

Lozier, R.H., Bogomolni, R.A., and Stoeckenius, W. (1975). Bacteriorhodopsin: a light-driven proton pump in *Halobacterium halobium*. Biophys. J. *15*, 955–962.

Martin, W., and Muller, M. (1998). The hydrogen hypothesis for the first eukaryote. Nature *392*, 37–41.

Matte-Tailliez, O., Brochier, C., Forterre, P., and Philippe, H. (2002). Archaeal phylogeny based on ribosomal proteins. Mol. Biol. Evol. *19*, 631–639.

Mevarech, M., and Werczberger, R. (1985). Genetic transfer in *Halobacterium volcanii*. J. Bacteriol. *162*, 461–462.

Mullakhanbhai, M.F., and Larsen, H. (1975). *Halobacterium volcanii* spec. nov., a Dead Sea halobacterium with a moderate salt requirement. Arch. Microbiol. *104*, 207–214.

Mylvaganam, S., and Dennis, P.P. (1992). Sequence heterogeneity between the two genes encoding 16S rRNA from the halophilic archaebacterium *Haloarcula marismortui*. Genetics *130*, 399–410.

Naor, A., Lapierre, P., Mevarech, M., Papke, R.T., and Gophna, U. (2012). Low species barriers in halophilic archaea and the formation of recombinant hybrids. Curr. Biol. *22*, 1444–1448.

Nelson-Sathi, S., Dagan, T., Landan, G., Janssen, A., Steel, M., McInerney, J.O., Deppenmeier, U., and Martin, W.F. (2012). Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. Proc. Natl. Acad. Sci. U.S.A. *109*, 20537–20542.

Ng, W.V., Kennedy, S.P., Mahairas, G.G., Berquist, B., Pan, M., Shukla, H.D., Lasky, S.R., Baliga, N.S., Thorsson, V., Sbrogna, J., *et al.* (2000). Genome sequence of *Halobacterium* species NRC-1. Proc. Natl. Acad. Sci. U.S.A. *97*, 12176–12181.

O'Malley, M.A. (2007). The nineteenth century roots of 'everything is everywhere'. Nat. Rev. Microbiol. *5*, 647–651.

Oh, D., Porter, K., Russ, B., Burns, D., and Dyall-Smith, M. (2010). Diversity of *Haloquadratum* and other haloarchaea in three, geographically distant, Australian saltern crystallizer ponds. Extremophiles *14*, 161–169.

Oren, A. (2008). Microbial life at high salt concentrations: phylogenetic and metabolic diversity. Saline Syst. *4*, 2.

Oren, A., Ventosa, A., and Grant, W.D. (1997). Proposed minimal standards for description of new taxa in the order Halobacteriales. Int. J. Syst. Bacteriol. *47*, 233–238.

Ortenberg, R., Tchelet, R., and Mevarech, M. (1999). A model for the genetic exchange system of the extremely halophilic archaeon *Haloferax volcanii*. In Microbiology and Biogenochemistry of Hypersaline Environments, Oren, A., ed. (CRC Press, Boca Raton), pp. 331–338.

Papke, R.T. (2009). A critique of prokaryotic species concepts. Methods Mol. Biol. *532*, 379–395.

Papke, R.T., and Gogarten, J.P. (2012). How bacterial lineages emerge. Science *336*, 45–46.

Papke, R.T., and Ward, D.M. (2004). The importance of physical isolation to microbial diversification. FEMS Microbiol. Ecol. *48*, 293–303.

Papke, R.T., Koenig, J.E., Rodriguez-Valera, F., and Doolittle, W.F. (2004). Frequent recombination in a saltern population of *Halorubrum*. Science *306*, 1928–1929.

Papke, R.T., Zhaxybayeva, O., Feil, E.J., Sommerfeld, K., Muise, D., and Doolittle, W.F. (2007). Searching for species in haloarchaea. Proc. Natl. Acad. Sci. U.S.A. *104*, 14092–14097.

Redfield, R.J. (2001). Do bacteria have sex? Nat. Rev. Genet. *2*, 634–639.

Rodriguez-Brito, B., Li, L., Wegley, L., Furlan, M., Angly, F., Breitbart, M., Buchanan, J., Desnues, C., Dinsdale, E., Edwards, R., *et al.* (2010). Viral and microbial community dynamics in four aquatic environments. ISME J. *4*, 739–751.

Rosenshine, I., Tchelet, R., and Mevarech, M. (1989). The mechanism of DNA transfer in the mating system of an archaebacterium. Science *245*, 1387–1389.

Sato, T., Fukui, T., Atomi, H., and Imanaka, T. (2003). Targeted gene disruption by homologous recombination in the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1. J. Bacteriol. *185*, 210–220.

Shapiro, B.J., Friedman, J., Cordero, O.X., Preheim, S.P., Timberlake, S.C., Szabo, G., Polz, M.F., and Alm, E.J. (2012). Population genomics of early events in the ecological differentiation of bacteria. Science *336*, 48–51.

Sharma, A.K., Spudich, J.L., and Doolittle, W.F. (2006). Microbial rhodopsins: functional versatility and genetic mobility. Trends Microbiol. *14*, 463–469.

Sharma, A., Walsh, D., Bapteste, E., Rodriguez-Valera, F., Ford Doolittle, W., and Papke, R.T. (2007). Evolution of rhodopsin ion pumps in haloarchaea. BMC Evol. Biol. *7*, 79.

Sheppard, S.K., McCarthy, N.D., Falush, D., and Maiden, M.C. (2008). Convergence of *Campylobacter* species: implications for bacterial evolution. Science *320*, 237–239.

Smith, H.O., Tomb, J.F., Dougherty, B.A., Fleischmann, R.D., and Venter, J.C. (1995). Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome. Science *269*, 538–540.

Smith, J.M., Feil, E.J., and Smith, N.H. (2000). Population structure and evolutionary dynamics of pathogenic bacteria. Bioessays *22*, 1115–1122.

Spudich, J.L., and Bogomolni, R.A. (1988). Sensory rhodopsins of Halobacteria. Annu. Rev. Biophys. Biophys. Chem. *17*, 193–215.

Stackebrandt, E., and Goebel, B.M. (1994). Taxonomic note: a place for DNA–DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int. J. Syst. Evol. Microbiol. *44*, 846–849.

Steinmoen, H., Knutsen, E., and Havarstein, L.S. (2002). Induction of natural competence in *Streptococcus pneumoniae* triggers lysis and DNA release from a subfraction of the cell population. Proc. Natl. Acad. Sci. U.S.A. *99*, 7681–7686.

Thomas, C.M., and Nielsen, K.M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat. Rev. Microbiol. *3*, 711–721.

Tripepi, M., Imam, S., and Pohlschroder, M. (2010). *Haloferax volcanii* flagella are required for motility but are not involved in PibD-dependent surface adhesion. J. Bacteriol. *192*, 3093–3102.

Vreeland, R.H., Straight, S., Krammes, J., Dougherty, K., Rosenzweig, W.D., and Kamekura, M. (2002). *Halosimplex carlsbadense* gen. nov., sp. nov., a unique halophilic archaeon, with three 16S rRNA genes, that grows only in defined medium with glycerol and acetate or pyruvate. Extremophiles *6*, 445–452.

Vulic, M., Dionisio, F., Taddei, F., and Radman, M. (1997). Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc. Natl. Acad. Sci. U.S.A. *94*, 9763–9767.

Wellner, A., Lurie, M.N., and Gophna, U. (2007). Complexity, connectivity, and duplicability as barriers to lateral gene transfer. Genome Biol. *8*, R156.

Whitaker, R.J., Grogan, D.W., and Taylor, J.W. (2005). Recombination shapes the natural population structure of the hyperthermophilic archaeon *Sulfolobus islandicus*. Mol. Biol. Evol. *22*, 2354–2361.

Williams, D., Gogarten, J.P., and Papke, R.T. (2012). Quantifying homologous replacement of loci between haloarchaeal species. Genome Biol. Evol. *4*, 1223–1244.

Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U.S.A. *74*, 5088–5090.

Woese, C.R., Kandler, O., and Wheelis, M.L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. U.S.A. *87*, 4576–4579.

Worrell, V.E., Nagle, D.P. Jr., McCarthy, D., and Eisenbraun, A. (1988). Genetic transformation system in the archaebacterium *Methanobacterium thermoautotrophicum* Marburg. J. Bacteriol. *170*, 653–656.

Yap, W.H., Zhang, Z., and Wang, Y. (1999). Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. J. Bacteriol. *181*, 5201–5209.

Zawadzki, P., Roberts, M.S., and Cohan, F.M. (1995). The log-linear relationship between sexual isolation and sequence divergence in *Bacillus* transformation is robust. Genetics *140*, 917–932.

Zhaxybayeva, O., Gogarten, J.P., Charlebois, R.L., Doolittle, W.F., and Papke, R.T. (2006). Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res. *16*, 1099–1108.

Zhaxybayeva, O., Doolittle, W.F., Papke, R.T., and Gogarten, J.P. (2009). Intertwined evolutionary histories of marine *Synechococcus* and *Prochloroccus marinus*. Genome Biol. Evol. *1*, 325–339.

Zhaxybayeva, O., Stepanauskas, R., Mohan, N.R., and Papke, R.T. (2013). Cell sorting analysis of geographically separated hypersaline environments. Extremophiles *17*, 265–275.

Chapter 4.2 Restriction-methylation genes in the Halobacteria

Introduction

The aims of this project are three-fold. The first is to identify the restriction-methylation system genes in the class Halobacteria. The second is to analyze the distribution and horizontal transfer of the identified genes. Finally, the ultimate goal is to explore the role of geography in speciation.

RM systems (RMS) have been selected for these investigations for several reasons. An RMS pairs a methylase (MTase) that targets a specific motif with a cognate endonuclease that targets the same motif (restriction enzyme or REase) (Roberts et al., 2003). This has lead to a longstanding hypothesis that RMS may provide an immunity function for the host (Kobayashi, 2001). The REase attacks un-methylated DNA, which presumably would be non-self DNA, and prevents incorporation of a large elements such as a virus into the genome (Corvaglia et al., 2010). Additionally, these systems are known to function in manner analogous to addiction cassettes (Ohno et al., 2008), and as such can also be viewed simply as selfish genetic elements making it difficult for the host to purge them from its genome. Regardless of which concept is correct, and they are not mutually incompatible, their functionality implies another possible role. They may play roles in initiating and/or propagating prokaryotic divergence (Budroni et al., 2011; Corvaglia et al., 2010; Kobayashi, 2001) by facilitating destruction of large stretches of interloping DNA. It thus becomes difficult for foreign ORFs to find their way into the host genome.

Methods

Search Approach. The starting data consists of 217 Halobacteria genomes from NCBI and 14 in-house sequenced genomes (**Table01)**. Queries for all restriction-methylation-specificity genes were obtained from the Restriction Enzyme dataBASE (REBASE) website (Roberts and Macelis, 2001). As methylation genes are classified by function rather than by homology (Roberts and Macelis, 2001) the protein sequences of each category were clustered into homologous groups (HGs) via the *uclust* function of the usearch (v9.0.2132) package (Edgar, 2010) at a 40% pid. The resulting ~36,000 HGs were aligned with MUSCLE v3.8.31 (Edgar, 2004). HMMs were then generated from the alignments using the *hmmbuild* function of HMMER3 v. 3.1b2 (hmmer.org). The orfs of the 217 genomes were searched against the profiles via the *hmmsearch* function of HMMER3. Top hits were extracted and cross hits filtered with in-house Perl scripts. Steps were taken to collapse and filter HGs. First, the hits were searched against the arCOG database (Makarova et al., 2015) using BLAST (Altschul et al., 1997) to assign arCOG identifiers to the members of each group. Second the R package *igraph* (Csardi and Nepusz, 2006) was used to create a list of connected components from the arCOG identifications. All members of a connected component were collapsed into a single collapsed HG (cHG).

Because REBASE is a database of all methylation-restriction-related activities there are many members of the database outside our interest. At this point we made a manual curation of our cHGs attempting to identify known functions that did not apply to our area of interest. Examples include protein methylation enzymes, exonucleases, cell-division proteins, etc. The final tally of this clustering and filtering yielded 1696 hits

across 48 total candidate cHGs. 26 cHGs are strong candidates with arCOG annotation suggesting DNA methylase activity, restriction enzyme activity, or specificity module activity as part of an RMS system. 22 are weaker candidates with predominant arCOG annotations matching other functions that may reasonably be excluded from conservative RMS-specific analyses. For a graphical representation of the search strategy see **Figure01**.

Reference phylogeny. A reference tree was created using the full complement of ribosomal proteins. The ribosomal protein set for *Halorubrum lacusprofundi* ATCC 49239 was obtained from the BioCyc website (Caspi et al., 2010). Each protein orf was used as the query in a BLAST (Altschul et al., 1997) search against each genome. Hits for each gene were aligned with MUSCLE v3.8.31 (Edgar, 2004) and then concatenated with in-house scripting. The concatenated alignment was subjected to maximum likelihood phylogenetic inference in the IQ-TREE v1.6.1 suite with ultrafast bootstrapping and automated model selection (Hoang et al.; Nguyen et al., 2015). The final model selection was LG+F+R9.

Presence-Absence Plot. The presence-absence matrix of cHGs was plotted as a heatmap onto the reference phylogeny using the *qheatmap* function of the R Bioconductor package *ggtree2* (Yu Guangchuang et al., 2016).

Rarefaction Curve of cHGs in Genomes. The rarefaction curve was generated with the *specaccum* function of the *vegan* package in R (Dixon, 2003).

Correlogram. Spearman correlations and significances between the presence-absence of cHGs was calculated with the *rcorr* function of the *hmisc* package in R. A significance cutoff of $p < 0.05$ was used with a bonferroni correction. All comparisons failing this criterion were set to correlation = 0. These data were plotted into a correlogram via the *corrplot* function of the R package *corrplot*.

Horizontal Gene Transfer Detection. Gene trees for each of the cHGs were inferred using RAxML v8.2.11 (Stamatakis, 2014) under PROTCATLG models with 100 bootstraps. The gene trees were then improved by resolving their poorly supported in nodes to match the species tree using TreeFix-DTL (Bansal et al., 2015). Optimized gene tree rootings were inferred with the OptRoot function of Ranger-DTL. Reconciliation costs for each gene tree were computed against the reference tree using Ranger-DTL 2.0 (http://compbio.engr.uconn.edu/software/RANGER-DTL/) (Bansal et al., 2012) with default DTL costs. One-hundred reconciliations, each using a different random seed, were calculated for each cHG. After aggregating these with the AggregateRanger function of Ranger-DTL the results were summarized and each prediction and any transfer inferred in 51% or greater of cases was counted as a transfer event.

Recognition Site Assignment. To determine the most likely recognition sites, each member of each cHG was searched against the REBASE Gold Standard set using

BLASTp. The REBASE gold standard set was chosen over the individual gene sets on account of it having a much higher density of recognition site annotation. This simplifies the need to search for secondary hits to find predicted target sites. After applying an e-value cutoff of 1E-20, the top hit was assigned to each ORF.

F81 Presence-absence Phylogeny. It is desirable to use maximum-likelihood methodology rather than simple distance measures. To realize this, the matrix was converted to an A/T alignment by replacing each present with an "A" and absent with a "T." This allowed use of an F81 model with empirical base frequencies. This confines the base parameters to only A and T while allowing all of the other advantages of an ML approach. IQ-TREE was employed to infer the tree with 100 bootstraps (Nguyen et al., 2015).

Internode Certainty. Tree certainty scores were calculated using the IC/TC score calculation algorithm implemented in RAxML v8.2.11 (Salichos and Rokas, 2013; Stamatakis, 2014).

Results.

The final tally of the homologous group clustering and filtering yielded 48 total candidate cHGs. 26 are strong candidates with arCOG annotation suggesting DNA methylase activity, restriction enzyme activity, or specificity module activity as part of an RMS system. 22 are weaker candidates with predominant arCOG annotations matching other functions are may reasonably be excluded from conservative RMS-specific analyses (**Table02)**. The majority of candidate RMS cHGs are present in fewer than half the genomes (**Figure02)**. Rarefaction analysis indicates all taxa are, on average, represented by half of the cHGs (**Figure03)**.

The phylogeny of the Halobacteria inferred from concatenation of ribosomal proteins was largely orthodox and broadly comparable to prior work (**Figure04)** (Soucy et al., 2014).

The distribution of RMS candidates throughout the Haloarchaea is highly patchy and does not appear to follow a clear pattern of vertical descent (**Figure05)**. This appearance was investigated by plotting the Jaccard distance of the presence-absence data against the alignment distance of the reference tree (**Figure06)**. If the presence-absence data followed vertical descent one would expect the best-fit line to follow a roughly 45-degree angle (a.k.a., a 1:1 relationship) or something close to it. Instead, the best fit line is essentially horizontal, indicating no significant relationship between the two variables.

To further evaluate the lack of vertical descent in the presence-absence pattern a phylogeny was inferred. The resultant tree is largely in disagreement with the reference

phylogeny (**Figure07)**. To further visualize on this point a tanglegram was constructed pairing the reference with presence-absence topologies (**Figure08)**. This view showcases the lack of similar topology between the two phylograms. The final point on this axis of enquiry is computing Internode Certainty scores for the reference tree using the support set from the F81 tree. The average IC score is an impressively low -0.509. IC is scaled from positive 1, when there is absolutely no conflicting signal in the support set to negative 1 when the support set supports alternative topologies.

The patchy distribution of the RMS candidates and their lack of conformity to the reference phylogeny suggests a large volume of horizontal gene transfer events as the most probable explanation for the observed data. To quantify the amount of transfer the TreeFix-Ranger pipeline was employed. TreeFix-DTL resolves poorly supported areas of gene trees to better match the species tree. Ranger-DTL resolves optimal gene tree rooting against the species tree and then computes a reconciliation estimating the number of duplications, transfers, and losses that best explains the data (**Table03)**. The pipeline reports a high volume of gene transfers in almost every cHG with four or more taxa. Approximately half (20 transfer events) of all (39 taxa) leaves in the average gene tree have experienced an HGT event. Only one cHG, a putative type III restriction module, has not been inferred to undergo at least one transfer event.

RMS systems usually function as cooperative units (Ohno et al., 2008; Roberts et al., 2003; Roberts and Macelis, 2001). It stands to reason that some of the RMS candidates may be transferred as units, maintaining their cognate functionality. This possibility was

examined by a correlation analysis. A spearman correlation was made between all pairs of cHGs. Those with a significant result at a Bonferroni-corrected $p < 0.05$ were plotted in a correlogram (**Figure09)**.

Discussion.

One of the striking points of these results is the irregular distribution of the RMS candidates throughout not just the class, but also within genera, species, and even communities. The patchy distribution is almost certainly the result of rampant HGT. While the sheer scale of the HGT is perhaps surprising, its existence as a major force is not. What really stands out is how little RMS genes *ever* seem to define a clade or an isolation source. *Halorubrum* only holds 5 candidate RMS cHGs absent from the remainder of the Halobacteria. And only one of those is found in more than 3 genomes, a type III restriction protein found in only 14 of 57 *Halorubrum* genomes. The *Halorubrum* distribution of presence and absences on the whole is only on the fringe of being different from the rest of the Halobacteria ($p = 0.04$, paired t-test). Essentially, HGT may be so prevalent that it might be that the presence of an RMS gene in one genome is equivalent to it being in any other haloarchaeal genome. While this author finds that conclusion a little extreme it is not entirely ridiculous. A culture-independent sampling of viruses from hypersaline environments pointed towards environments that are thousands of kilometers away from each other sharing a common 'hypersaline-ness' in their common viral assemblages (Santos et al., 2012). An examination of viruses from sites in Italy and Thailand found viruses with very similar genome sequences and morphological structure (Senčilo et al., 2013). Finally, a recent study examining halophile-virus interactions from

samples across the globe reported that viral host ranges frequently crossed great geographic range (Atanasova et al., 2012). The implications being support “for the idea that there is a global exchange of microbes and their viruses.” And also that “It suggests that hypersaline environments worldwide function like a single habitat.” (Atanasova et al., 2012). This is a little reminiscent of some of the early ideas of cellular life (Kandler, 2002; Lawrence, 1999; Woese, 1998). It also harkens to, and exceeds, the level of genome fluidity for which the pan-genome concept was created to explain (Lapierre and Gogarten, 2009; Tettelin et al., 2005). Finally, this author sees a pleasing resonance with the Strong Black Queen concept (Fullmer et al., 2015). Still, if the hypothesis of RMS genes existing to act a defense against foreign selfish genetic elements is correct, how can a lineage get by without a robust defense?

Since one of the primary targets is viral intrusion (Furuta and Kobayashi, 2013) perhaps there is a plausible explanation for why RMS is not an essential part of cellular countermeasures. The first consideration needs to be whether viruses are a threat strong enough to select for strong defenses. As predation of prokaryotes is very rare in high salt concentration, on account of Eukaryotes finding the environs inhospitable, viruses are generally seen as the main biological factor controlling prokaryotic populations (Guixa-Boixareu et al., 1996). In support of this idea microscopy reports that the number of virus particles in relation to the number of cells increases as salinity rises (Santos et al., 2012). Seemingly in opposition, lysis is inducible in some viruses by lowering the salinity (Santos et al., 2012). Other work has also suggested that as salinity moves towards 30% viruses tend to behave more temperately (Bettarel et al., 2011). It may be that viruses are

less of an acute threat to halophiles than is typical, however, the disproportionate particle counts argue they must still be a major source of selection.

If the virions are allowed to dock and import then other defenses besides RMS may be ready to resist. The best known of these defenses is the CRISPR-Cas system (Gophna and Brodt, 2012). CRISPR recognizes short (~40bp) regions of invading DNA that the host has been exposed to previously and degrades it. While it appears to be very successful at its task, prior work has indicated that it is not cosmopolitan within communities and close phylogenetic clusters, such as the denizens of the Aran-Bidgol lake (Fullmer et al., 2014). That leaves preventing phage from infiltrating the cell as the most likely defense. Altering surface decoration is one of the primary methods of avoiding phage predation. In the *Haloferax* there are two pathways which control glycosylation of external features. One is relatively stable while the other is highly variable and shows hallmarks of having genes mobilized for horizontal transfer (Shalev et al., 2018). On this thread, at least one halovirus has been found to require glycosylation by its host in order to infect properly (Kandiba et al., 2012). Another study found widespread "metagenomic" islands in closely related genomes of halophilic prokaryotes (Rodriguez-Valera et al., 2009). Each of these islands is filled with a seemingly unique mixture of LPS and other external structure genes.

Does any of this explain why RMS candidates are not found in cosmopolitan fixation? Probably not. Perhaps the constant fight for advantage proposed under the constant-diversity model (Rodriguez-Valera et al., 2009) is allowing rare genotypes lacking

intracellular defenses to survive. At least temporarily, until they become the leaders in the population and the viruses shift to attack them. It could be that as this happens they can then acquire RM immunity through the apparently ubiquitous HGT and the lineage is saved. Or maybe they do not and the lineage is slaughtered until the viruses move on to another upstart collection of genotypes that lost their defenses in exchange for a growth benefit. Another alternative could be that viruses, or other infiltrating selfish genetic elements, might gain access to the host's methylation after any successful infection that is not stopped by the restriction system. In that case, a limited and vertically inherited RM system would then be an ineffective defense against the virus going forward. Under this scenario, a large and diverse pool of mobilized RMS genes could offer a stronger defense for the population as a whole. A single successful infection would no longer endanger the entire group of potential hosts. This dynamic may represent a form of balancing selection where compromised defenses are selected against until another scheme grows to prominence and is compromised.

Further work.

This work is far from complete and dozens of avenues of enquiry remain open. The earliest plan was to look at geographic versus phylogenetic relationships in gene content and function in the *Halorubrum*. That line remains almost as unanswered as the day this work began. An analysis examining the co-localization of RMS candidates in genomes is half-completed on the cluster as of this writing. It might provide much insight into whether these genes are being transferred as individual ORFs or as part of larger units or operons. Using PFAM to annotate domains in the cHGs may offer another method to

evaluate the relevance of the groups and their likely functions. While the existence of HGT has been demonstrated it would be interesting to try and determine how these genes are mobilized. One could survey the surrounding genomic area for insertions sequences, phage proteins, transposes, *tra* genes etc. for clues as to the mobilization. Presumably, any that are known to have been transferred but do not have a mechanism attached have been moved via the remarkable mating phenomenon. A recently discovered innate immunity system, BREX (BREX is a novel phage resistance system widespread in microbial genomes), has not yet been examined in this study and might offer some additional perspective on how this organisms, and particularly the community in Aran-Bidgol defend themselves from the viral threat.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi:10.1093/nar/25.17.3389.

Atanasova, N. S., Roine, E., Oren, A., Bamford, D. H., and Oksanen, H. M. (2012). Global network of specific virus–host interactions in hypersaline environments. *Environ. Microbiol.* 14, 426–440. doi:10.1111/j.1462-2920.2011.02603.x.

Bansal, M. S., Alm, E. J., and Kellis, M. (2012). Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics* 28, i283–i291. doi:10.1093/bioinformatics/bts225.

Bansal, M. S., Wu, Y.-C., Alm, E. J., and Kellis, M. (2015). Improved gene tree error correction in the presence of horizontal gene transfer. *Bioinformatics* 31, 1211–1218. doi:10.1093/bioinformatics/btu806.

Bettarel, Y., Thierry, B., Corinne, B., Claire, C., Anne, D., Isabelle, D., et al. (2011). Ecological traits of planktonic viruses and prokaryotes along a full-salinity gradient. *FEMS Microbiol. Ecol.* 76, 360–372. doi:10.1111/j.1574-6941.2011.01054.x.

BREX is a novel phage resistance system widespread in microbial genomes Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4337064/ [Accessed May 3, 2018].

Budroni, S., Siena, E., Dunning Hotopp, J. C., Seib, K. L., Serruto, D., Nofroni, C., et al. (2011). Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination. *Proc. Natl. Acad. Sci. U. S. A.* 108, 4494–4499. doi:10.1073/pnas.1019751108.

Caspi, R., Altman, T., Dale, J. M., Dreher, K., Fulcher, C. A., Gilham, F., et al. (2010). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 38, D473–D479. doi:10.1093/nar/gkp875.

Corvaglia, A. R., François, P., Hernandez, D., Perron, K., Linder, P., and Schrenzel, J. (2010). A type III-like restriction endonuclease functions as a major barrier to horizontal gene transfer in clinical Staphylococcus aureus strains. *Proc. Natl. Acad. Sci. U. S. A.* 107, 11954–11958. doi:10.1073/pnas.1000489107.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal* Complex Systems, 1695.

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* 14, 927–930. doi:10.1111/j.1654-1103.2003.tb02228.x.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi:10.1093/nar/gkh340.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi:10.1093/bioinformatics/btq461.

Fullmer, M. S., Soucy, S. M., and Gogarten, J. P. (2015). The pan-genome as a shared genomic resource: mutual cheating, cooperation and the black queen hypothesis. *Evol. Genomic Microbiol.*, 728. doi:10.3389/fmicb.2015.00728.

Fullmer, M. S., Soucy, S. M., Swithers, K. S., Makkay, A. M., Wheeler, R., Ventosa, A., et al. (2014). Population and genomic analysis of the genus Halorubrum. *Extreme Microbiol.* 5, 140. doi:10.3389/fmicb.2014.00140.

Furuta, Y., and Kobayashi, I. (2013). *Restriction-Modification Systems as Mobile Epigenetic Elements*. Landes Bioscience Available at: https://www.ncbi.nlm.nih.gov/books/NBK63963/ [Accessed December 5, 2017].

Gophna, U., and Brodt, A. (2012). CRISPR/Cas systems in archaea. *Mob. Genet. Elem.* 2, 63–64. doi:10.4161/mge.19907.

Guixa-Boixareu, N., Ji, C.-P., M, H., G, B., and C, P.-A. (1996). Viral lysis and bacterivory as prokaryotic loss factors along a salinity gradient. *Aquat. Microb. Ecol.* 11, 215–227. doi:10.3354/ame011215.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Le, S. V. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* doi:10.1093/molbev/msx281.

Kandiba, L., Aitio, O., Helin, J., Guan, Z., Permi, P., Bamford, D. H., et al. (2012). Diversity in prokaryotic glycosylation: an archaeal-derived N-linked glycan contains legionaminic acid. *Mol. Microbiol.* 84, 578–593. doi:10.1111/j.1365-2958.2012.08045.x.

Kandler, O. (2002). *The Early Diversification of Life and the Origin of the Three Domains: A Proposal*. CRC Press.

Kobayashi, I. (2001). Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res.* 29, 3742–3756.

Lapierre, P., and Gogarten, J. P. (2009). Estimating the size of the bacterial pan-genome. *Trends Genet.* 25, 107–110. doi:10.1016/j.tig.2008.12.004.

Lawrence, J. G. (1999). *Gene Transfer and Minimal Genome Size*. National Academies Press (US) Available at: https://www.ncbi.nlm.nih.gov/books/NBK224755/ [Accessed June 12, 2018].

Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2015). Archaeal Clusters of Orthologous Genes (arCOGs): An Update and Application for Analysis of Shared Features between Thermococcales, Methanococcales, and Methanobacteriales. *Life Basel Switz.* 5, 818–840. doi:10.3390/life5010818.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi:10.1093/molbev/msu300.

Ohno, S., Handa, N., Watanabe-Matsui, M., Takahashi, N., and Kobayashi, I. (2008). Maintenance Forced by a Restriction-Modification System Can Be Modulated by a Region in Its Modification Enzyme Not Essential for Methyltransferase Activity. *J. Bacteriol.* 190, 2039–2049. doi:10.1128/JB.01319-07.

Roberts, R. J., Belfort, M., Bestor, T., Bhagwat, A. S., Bickle, T. A., Bitinaite, J., et al. (2003). A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.* 31, 1805–1812. doi:10.1093/nar/gkg274.

Roberts, R. J., and Macelis, D. (2001). REBASE—restriction enzymes and methylases. *Nucleic Acids Res.* 29, 268–269. doi:10.1093/nar/29.1.268.

Rodriguez-Valera, F., Martin-Cuadrado, A.-B., Rodriguez-Brito, B., Pašić, L., Thingstad, T. F., Rohwer, F., et al. (2009). Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol.* 7, 828–836. doi:10.1038/nrmicro2235.

Salichos, L., and Rokas, A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497, 327–331. doi:10.1038/nature12130.

Santos, F., Yarza, P., Parro, V., Meseguer, I., Rosselló-Móra, R., and Antón, J. (2012). Viruses from hypersaline environments: a culture-independent approach. *Appl. Environ. Microbiol.*, AEM.07175-11. doi:10.1128/AEM.07175-11.

Senčilo, A., Jacobs-Sera, D., Russell, D. A., Ko, C.-C., Bowman, C. A., Atanasova, N. S., et al. (2013). Snapshot of haloarchaeal tailed virus genomes. *RNA Biol.* 10, 803–816. doi:10.4161/rna.24045.

Shalev, Y., Soucy, S. M., Papke, R. T., Gogarten, J. P., Eichler, J., and Gophna, U. (2018). Comparative Analysis of Surface Layer Glycoproteins and Genes Involved in Protein Glycosylation in the Genus Haloferax. *Genes* 9, 172. doi:10.3390/genes9030172.

Soucy, S. M., Fullmer, M. S., Papke, R. T., and Gogarten, J. P. (2014). Inteins as indicators of gene flow in the halobacteria. *Extreme Microbiol.* 5, 299. doi:10.3389/fmicb.2014.00299.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi:10.1093/bioinformatics/btu033.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome." *Proc. Natl. Acad. Sci. U. S. A.* 102, 13950–13955. doi:10.1073/pnas.0506758102.

Woese, C. (1998). The universal ancestor. *Proc. Natl. Acad. Sci.* 95, 6854–6859. doi:10.1073/pnas.95.12.6854.

Yu Guangchuang, Smith David K., Zhu Huachen, Guan Yi, Lam Tommy Tsan-Yuk, and McInerny Greg (2016). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* 8, 28–36. doi:10.1111/2041-210X.12628.

4.2.1 RMS Figures & Tables.

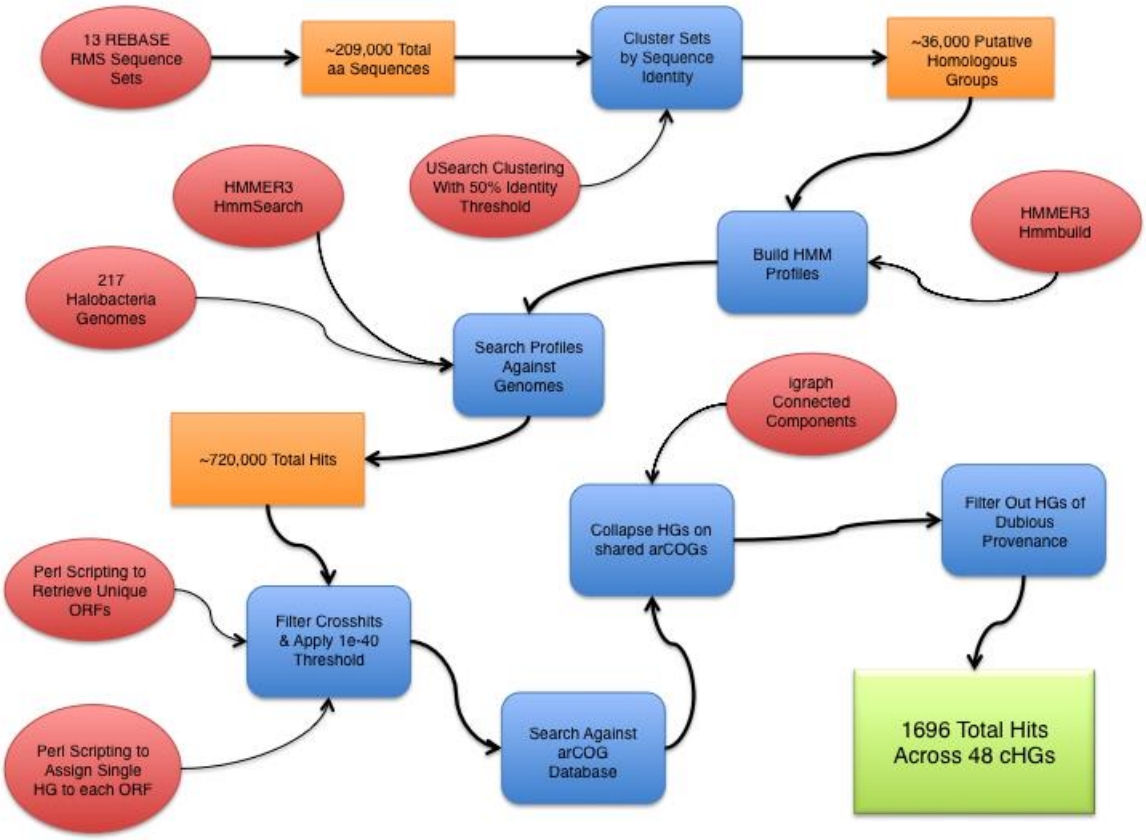


Figure01. Workflow of RMS-candidate gene search strategy. Red circles represent input data and tools used to operate on data. Blue rectangles represent use and modification of the data during the process. Orange rectangles represent outputs at significant intermediate points in the process.

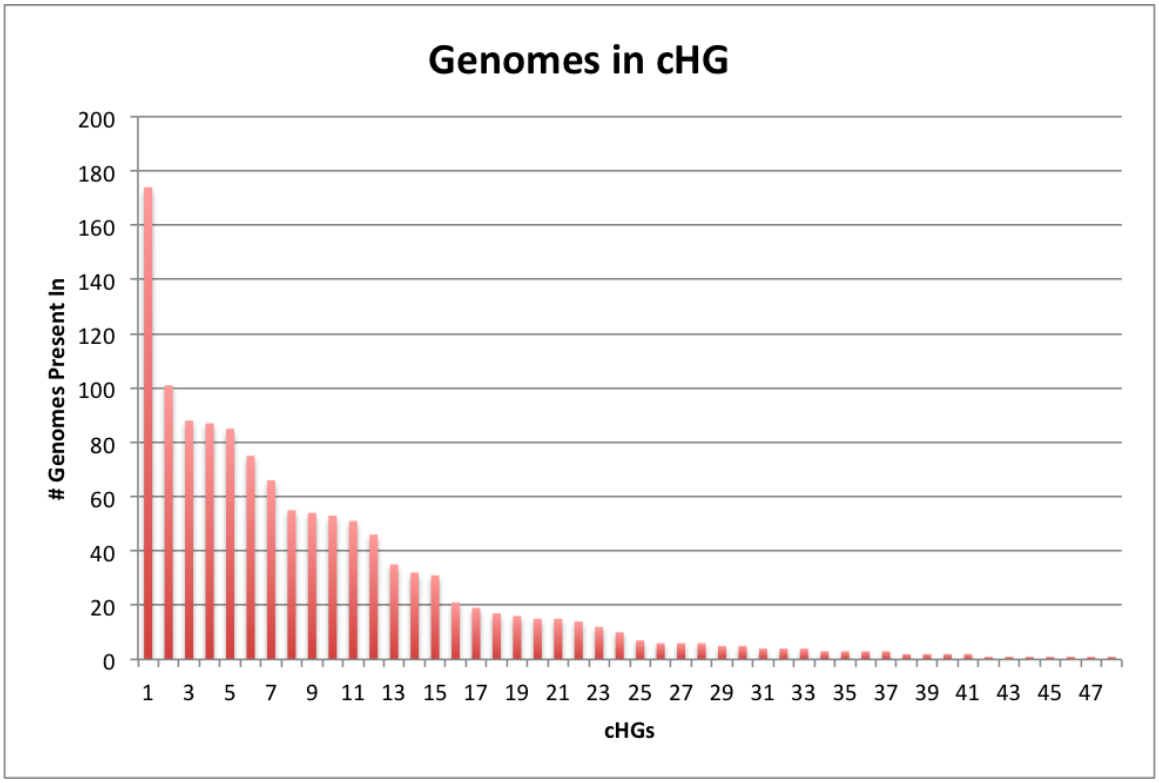


Figure02. Bargraph of the number of genomes present in each cHG. No cHG contains a representative from every genome used in this study. Indeed, all but one cHG contain members from fewer than half of the genomes.

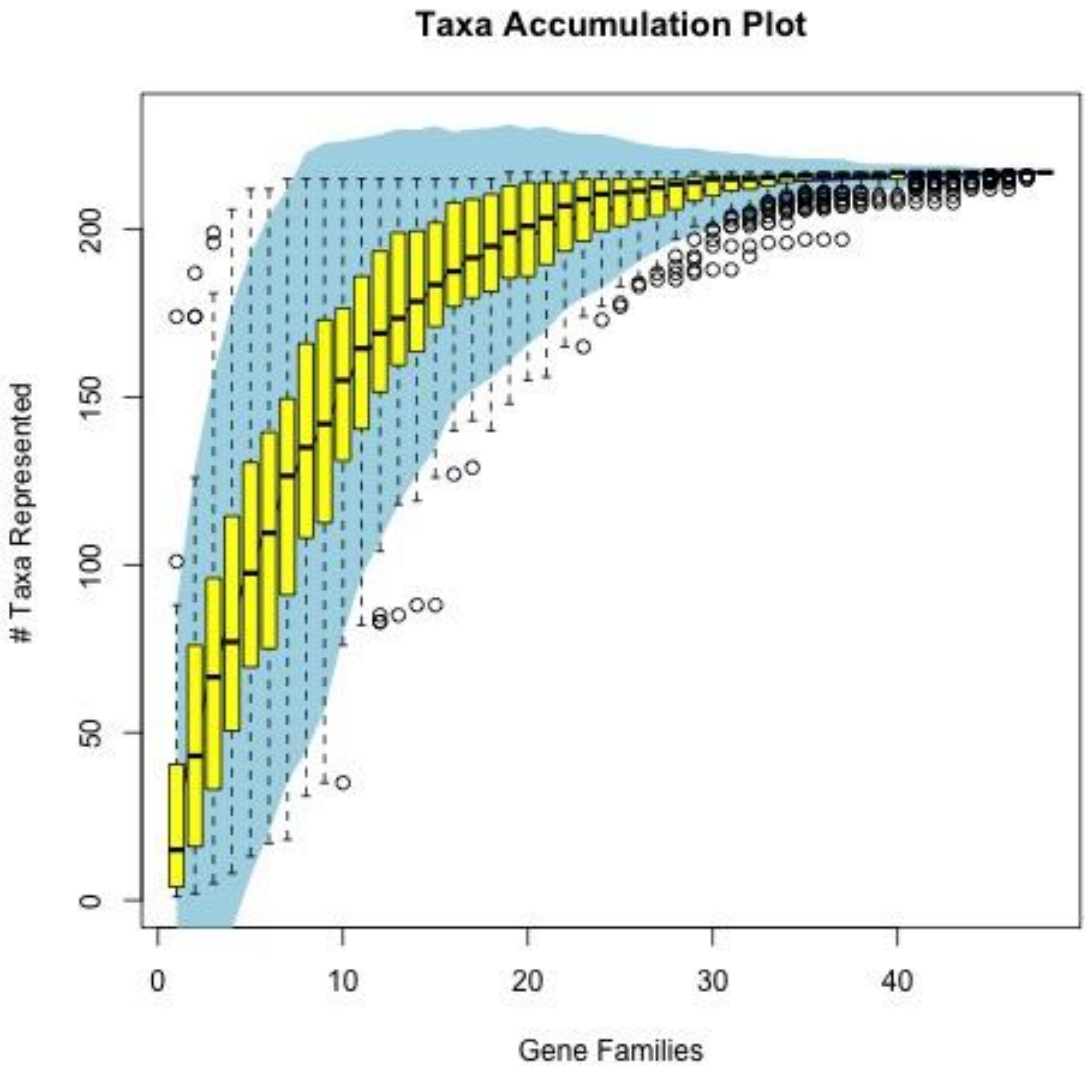


Figure03. Rarefaction plot of the number of genomes represented as cHGs accumulate.

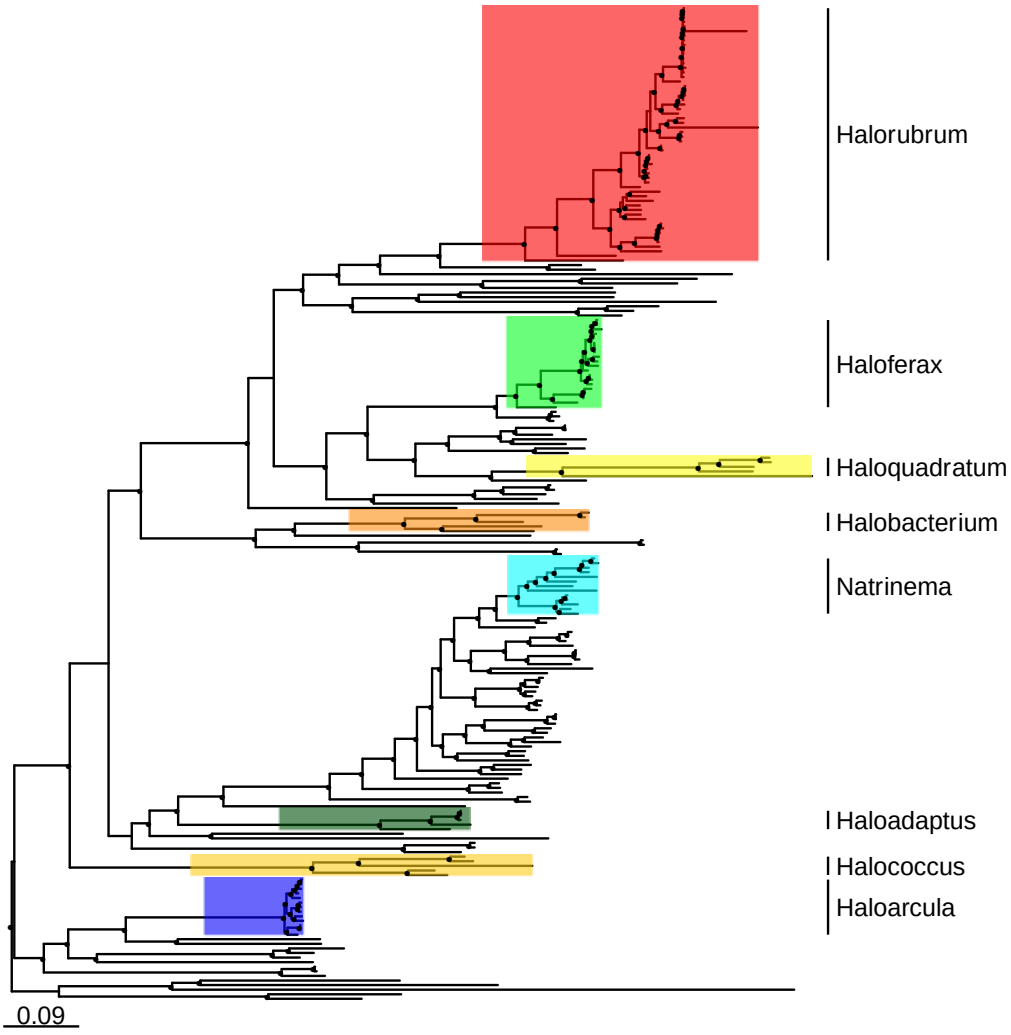


Figure04. Reference phylogeny inferred from a concatenation of ribosomal proteins. Tree was inferred under the LG substitution matrix using empirical base frequencies (+F) and a FreeRate (Soubrier et al., 2012) model of rate heterogeneity with nine categories (+R9). The major groups assemble as expected and the topology largely is in conformance with prior analysis (Soucy et al., 2014). The root was placed using Soucy et al., 2014 as guide. Dots represent bootstrap values greater than or equal to 80%.

A	T_I_M-021	Q	Adenine_DNA_methylase_probable_T_III_M-042	AG	Endonuclease-020
B	T_I_M-024	R	T_III_R-008	AH	HNH_endonuclease-004
C	T_I_R-018	S	T_III_R_probable-009	AI	HNH_nuclease-037
D	T_I_R-034	T	Adenine_DNA_methylase-014	AJ	HNH_nuclease-039
E	T_I_R-045	U	DNA_methylase-022	AK	HNH_nuclease-041
F	T_I_S-006	V	DNA_methylase-027	AL	MBF1-046
G	T_I_S-025	W	dam_methylase-031	AM	CBS_domain-028
H	probable_T_II_M-036	X	probable_RMS_M-035	AN	MarR-005
I	T_II_M-001	Y	dcm_methylase-044	AO	ParB-like_nuclease-030
J	T_II_M-003	Z	Adenine_DNA_methylase-048	AP	GVPC-016
K	T_II_M-011	AA	RNA_methylase-010	AQ	ASCH_domain_RNA-binding-002
L	T_II_M-033	AB	SAM-methylase-040	AR	Uncharacterized-017
M	T_II_R-007	AC	RestrictionEndonuclease-012	AS	Uncharacterized-026
N	T_II_R-013	AD	PredictedRestrictionEndonuclease-038	AT	Uncharacterized-032
O	T_II_R-023	AE	HNH_endonuclease-015	AU	Uncharacterized-043
P	T_II_R-029	AF	Endonuclease-019	AV	Uncharacterized-047

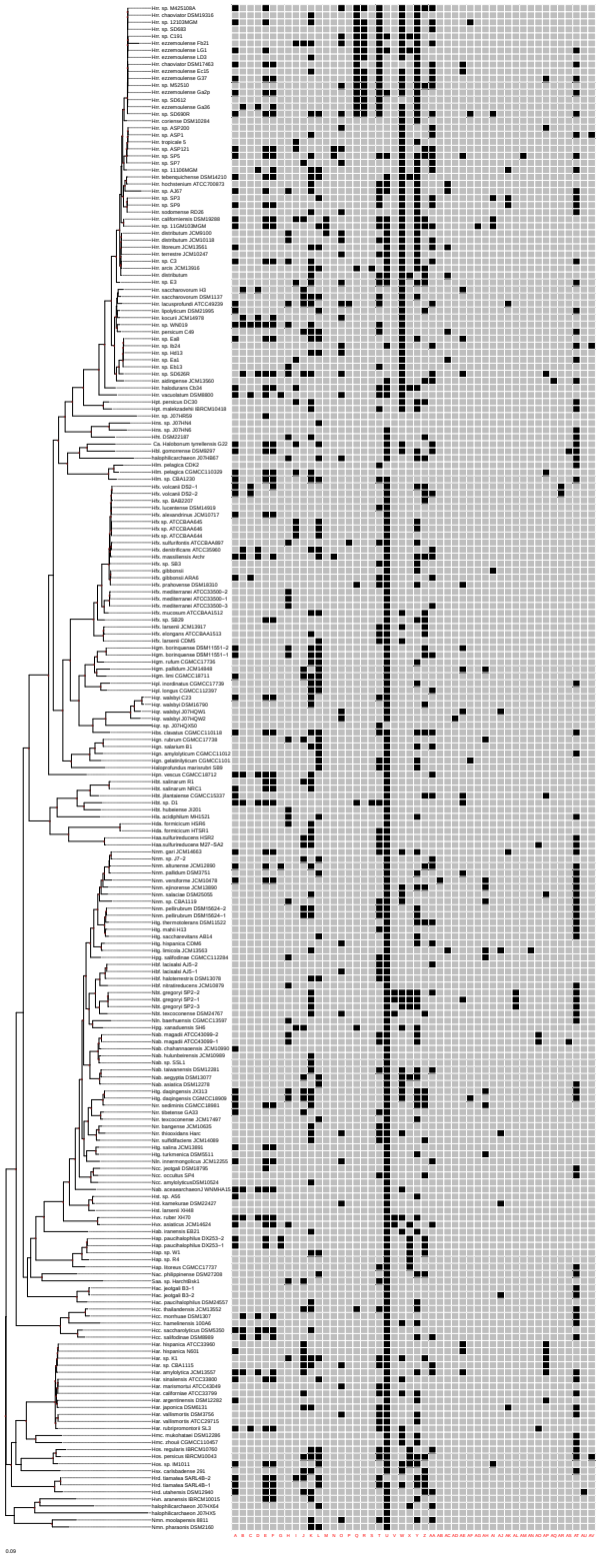

Hor. sp. M425106A
Hor. chaovator DSM19316
Hor. sp. 121024GM
Hor. sp. SD683
Hor. sp. C191
Hor. ezzemoulense Fb21
Hor. ezzemoulense LG1
Hor. ezzemoulense LD2
Hor. chaovator DSM17463
Hor. ezzemoulense Ec15
Hor. ezzemoulense G37
Hor. sp. MS2510
Hor. ezzemoulense Ga2p
Hor. sp. SD612
Hor. ezzemoulense Ga26
Hor. sp. SD680R
Hor. coreense DSM10284
Hor. sp. ASP200
Hor. sp. ASP1
Hor. tropicale 5
Hor. sp. ASP121
Hor. sp. SP5
Hor. sp. SP7
Hor. sp. 111064GM
Hor. lebanqiuchense DSM14210
Hor. hochstetterum ATCC700873
Hor. sp. AJ67
Hor. sp. SP3
Hor. sp. SP9
Hor. sodomense RD36
Hor. californiensis DSM19288
Hor. sp. 11GM103MGM
Hor. distributum JCM9100
Hor. distributum JCM10118
Hor. litoreum JCM13561
Hor. terrestre JCM10247
Hor. sp. C3
Hor. arcis JCM13916
Hor. distributum
Hor. sp. E3
Hor. saccharovorum H2
Hor. saccharovorum DSM1137
Hor. tacuapiruhandi ATCC49229
Hor. lipolyticum DSM21995
Hor. kocuri JCM14978
Hor. sp. WN019
Hor. persicum C49
Hor. sp. Ea8
Hor. sp. Bi24
Hor. sp. Hd13
Hor. sp. Ea1
Hor. sp. Eb13
Hor. sp. SD6269
Hor. salinigenae JCM13560
Hor. halobiaterni Ch34
Hor. vacuolatum DSM6800
Hpt. persicus DC30
Hpt. makkoyaderhi IBRCM10418
Hor. sp. JO74R59
Hms. sp. JO7HN4
Hms. sp. JO7HN6
Hht. DSM22187
Ca. Halobiterum tyrrellensis G22
Hbt. gomorrense DSM8297
halophilicarchaeon JO74B87
Hbm. pelagica CDK2
Hbm. pelagica CGMCC110329
Hbm. sp. CBA1230
Hlb. volcanii DS2–1
Hlb. volcanii DS2–2
Hlb. sp. BAB2207
Hlb. lucentense DSM14919
Hlb. alexandrinus JCM10717
Hlb sp. ATCCBAA645
Hlb sp. ATCCBAA646
Hlb sp. ATCCBAA644
Hlb. sulfurifontis ATCCBAA897
Hlb. denitrificans ATCC35960
Hlb. massiliensis Ache
Hlb. sp. SB3
Hlb. gibbonsi
Hlb. gibbonsi ARA6
Hlb. jeotgalense DSM18310
Hlb. mediterranei ATCC33500–2
Hlb. mediterranei ATCC33500–1
Hlb. mediterranei ATCC33500–3
Hlb. mucosum ATCCBAA1512
Hlb. sp. SB39
Hlb. terrenii JCM13917
Hlb. elongans ATCCBAA1513
Hlb. terrenii CSM5
Hgm. borinquense DSM11501–2
Hgm. borinquense DSM11501–4
Hgm. rufum CGMCC17736
Hgm. pallidum JCM14848
Hgm. limi CGMCC18711
Hgl. inordinatus CGMCC17739
Hgl. longus CGMCC110397
Hsp. walsbyi C23
Hsp. walsbyi DSM16790
Hsp. walsbyi JO74QW0
Hsp. walsbyi JO74QW2
Hsp. sp. JO74QX0
Hbs. clavatus CGMCC10116
Hgm. rubrum CGMCC17738
Hgm. salarium B1
Hgm. amylolyticum CGMCC11012
Hgm. gelatinilyticum CGMCC11201
Haloprofundus marinisolii S89
Hpm. vescus CGMCC18712
Hlb. salinarui R1
Hlb. salinarum NRC1
Hlb. jilantaiense CGMCC15322
Hbt. sp. D1
Hlb. habaiense J201
Hla. acidiphilum MH2521
Hda. formicicum HSR6
Hda. formicicum HTSR1
Haa.sulfurireducens HGR2
Haa.sulfurireducens MZF–SA2
Nmm. gari JCM14663
Nmm. sp. JT–2
Nmm. alkaliense JCM12890
Nmm. pallidum DSM3751
Nmm. vensiformi JCM10478
Nmm. ejinorense JCM18890
Nmm. salaciae DSM25055
Nmm. sp. CBA1119
Nmm. pellirubrum DSM15624–2
Nmm. pellirubrum DSM15624–1
Hlg. thermotolerans DSM11522
Hlg. mukhi H13
Hlg. saccharovitens AB14
Hlg. hispanica CSM6
Hlg. limicola JCM13563
Hgg. salifodinae CGMCC112284
Hbf. lacisalsi AJ5–2
Hbf. lacisalsi AJ5–1
Hbf. halobiterensis DSM13078
Hbf. nitratireducens JCM10879
Nbt. gregoryi SP2–2
Nbt. gregoryi SP2–1
Nbt. gregoryi SP2–3
Nth. teccoconense DSM24767
Nth. baerhamnis CGMCC13597
Hpg. xanaduensis SH6
Nab. magadii ATCC43099–2
Nab. magadii ATCC43099–1
Nab. chahannaoensis JCM10990
Nab. hubeiensis JCM10989
Nab. sp. SSL1
Nab. taiwanensis DSM12281
Nab. aegyptia DSM13077
Nab. asiatica DSM12278
Hlg. daqingensis JX313
Hlg. daqingensis CGMCC18909
Nor. sedimins CGMCC18981
Nor. tibetense GA33
Nor. teccoconense JCM17497
Nor. bangense JCM10635
Nor. thiooxidans Harc
Nm. sulfidifaciens JCM14089
Hlg. salina JCM13891
Hlg. turkmenica DSM5511
Nth. innermongolicus JCM12255
Ncc. jeotgali DSM18795
Ncc. occulbis SP4
Ncc. amylolyticus DSM10524
Nab. aceaearchaeus3 WNAH4-15
Hst. sp. A56
Hst. kamekurae DSM22427
Hst. tarsenii XH48
Hox. ruber XH70
Hox. asiaticus JCM14624
Hab. inarensis EB21
Hap. paucihalophilus DX253–2
Hap. paucihalophilus DX253–1
Hap. sp. W1
Hap. sp. R4
Hap. litoreus CGMCC17737
Nac. pHlippinense DSM27208
Saa. sp. Hasc08A43
Hac. jeotgali B3–1
Hac. jeotgali B3–2
Hac. pauchalophilus DSM24527
Hcc. thailandensis JCM13012
Hcc. mormuae DSM1307
Hcc. hamelinensis 16SA6
Hcc. saccharolyticus DSM5350
Hcc. salifodinae DSM48669
Hac. hispanica ATCC33960
Hac. hispanica N691
Har. sp. K1
Har. sp. CBA1115
Har. amylolytica JCM13557
Har. straleensis ATCC13830
Har. marismortui ATCC43049
Har. californiae ATCC33799
Har. argentinensis DSM12282
Har. japonica DSM6131
Har. vallismortis DSM3756
Har. vallismortis ATCC29715
Har. rubripromontori SL3
Hmc. mukohataei DSM12286
Hmc. zhouii CGMCC10467
Hos. regularis IBRCM10760
Hos. persicus IBRCM10043
Hos. sp. IM1011
Hos. carlsbadense 291
Hrd. litamalea SARL48–2
Hrd. litamalea SARL48–1
Hrd. utahensis DSM12940
Hun. aranensis IBRCM10015
halophilicarchaeon JO74X64
halophilicarchaeon JO74XI5
Nmm. mosdagensis BB11
Nmm. pharaonis DSM2160
Absent
Present
0.09

Figure05. Presence-absence matrix of the 48 candidate RMS cHGs plotted against the reference phylogeny. The pattern of presence-absence does not appear to match the reference phylogeny. RMS-candidate cHGs are loosely ordered by system type and with the dubious candidates at the end. Below is a also a key relating the column names to the majority functional annotation.

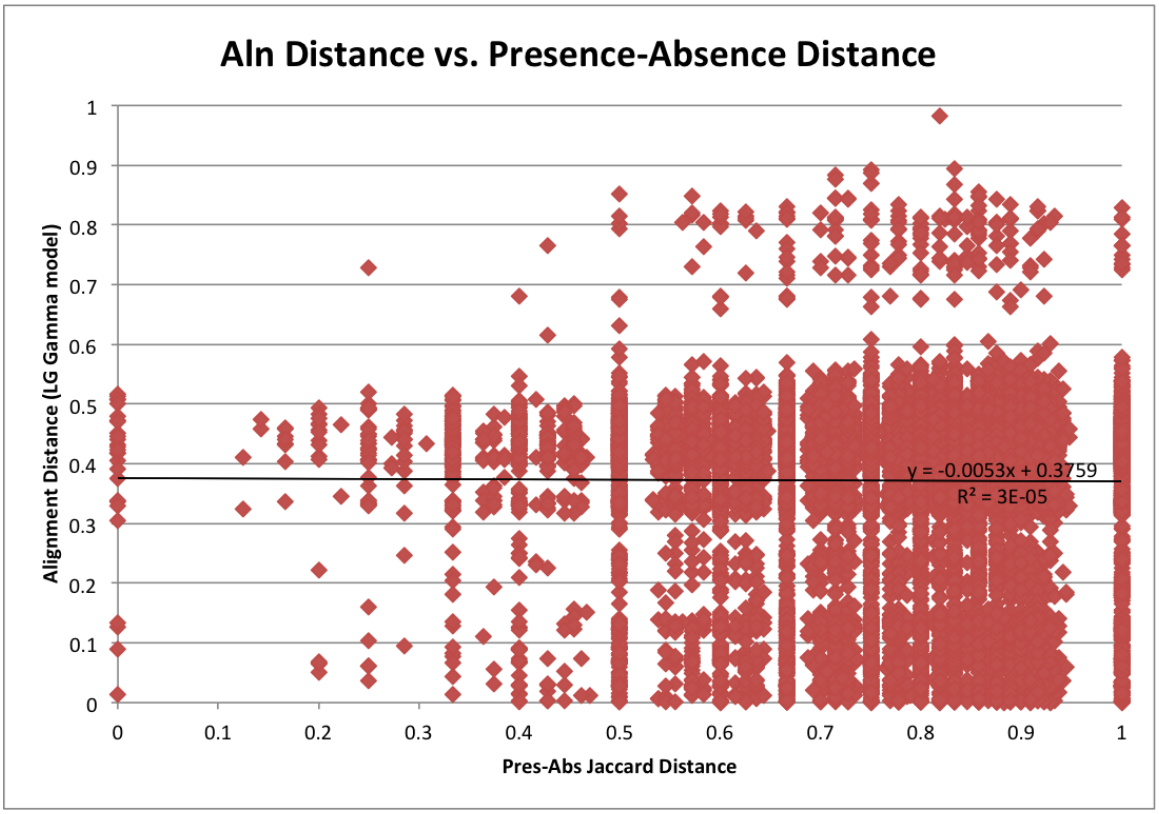


Figure06. Plot of alignment distance as a function of presence-absence distance. Alignment distance was calculated from an LG+Gamma model and presence-absence using jaccard distance. If presence-absence pattern is a function of vertical descent then the best fit line should broadly follow a 45-degree angle. There is clearly no correlation between the two measures.

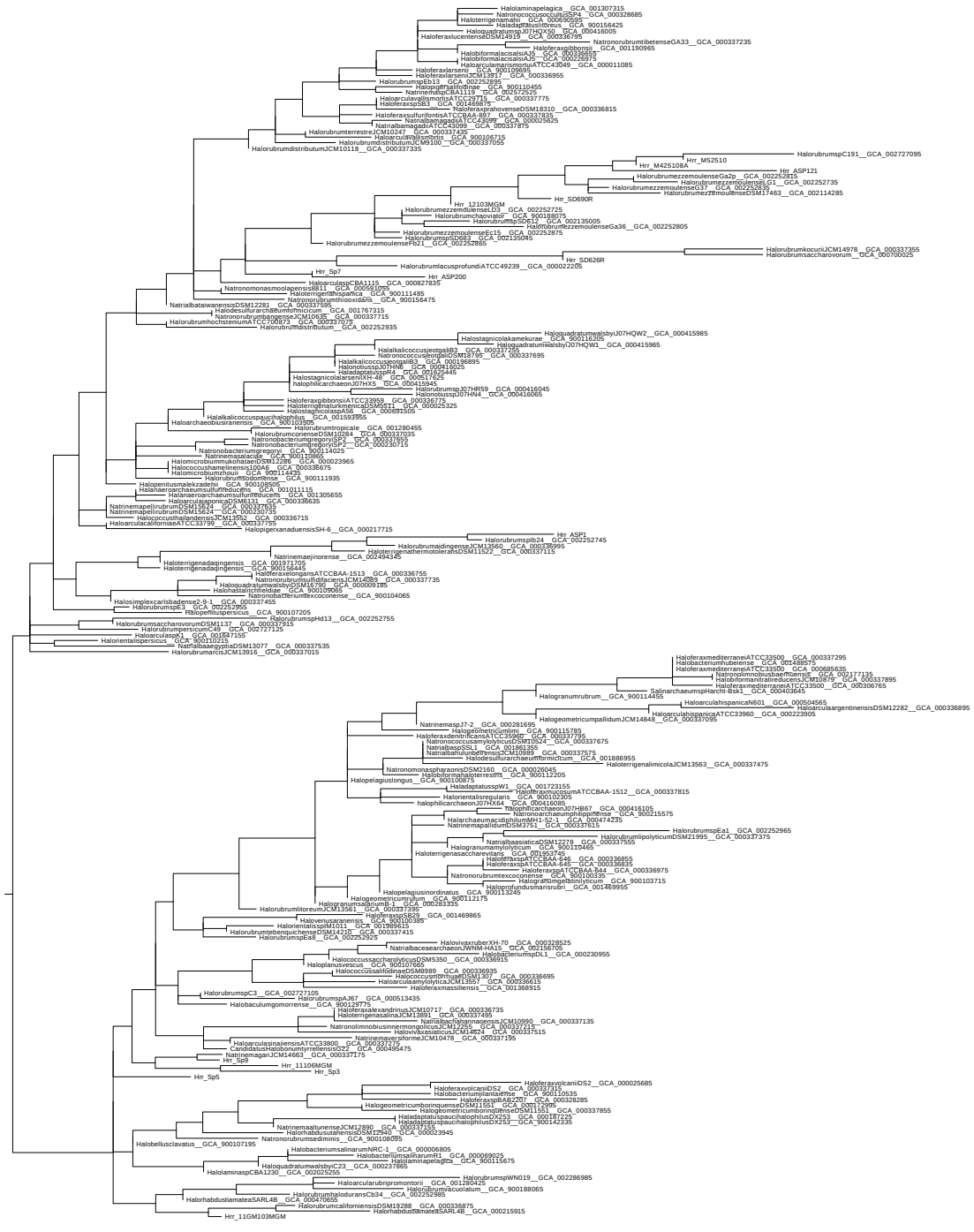


Figure07. Maximum-likelihood phylogeny of cHG presence-absence matrix. An F81 model with empirical base-frequency was employed. Root was placed at midpoint on account of the topology not bearing meaningful resemblance to the reference topology.

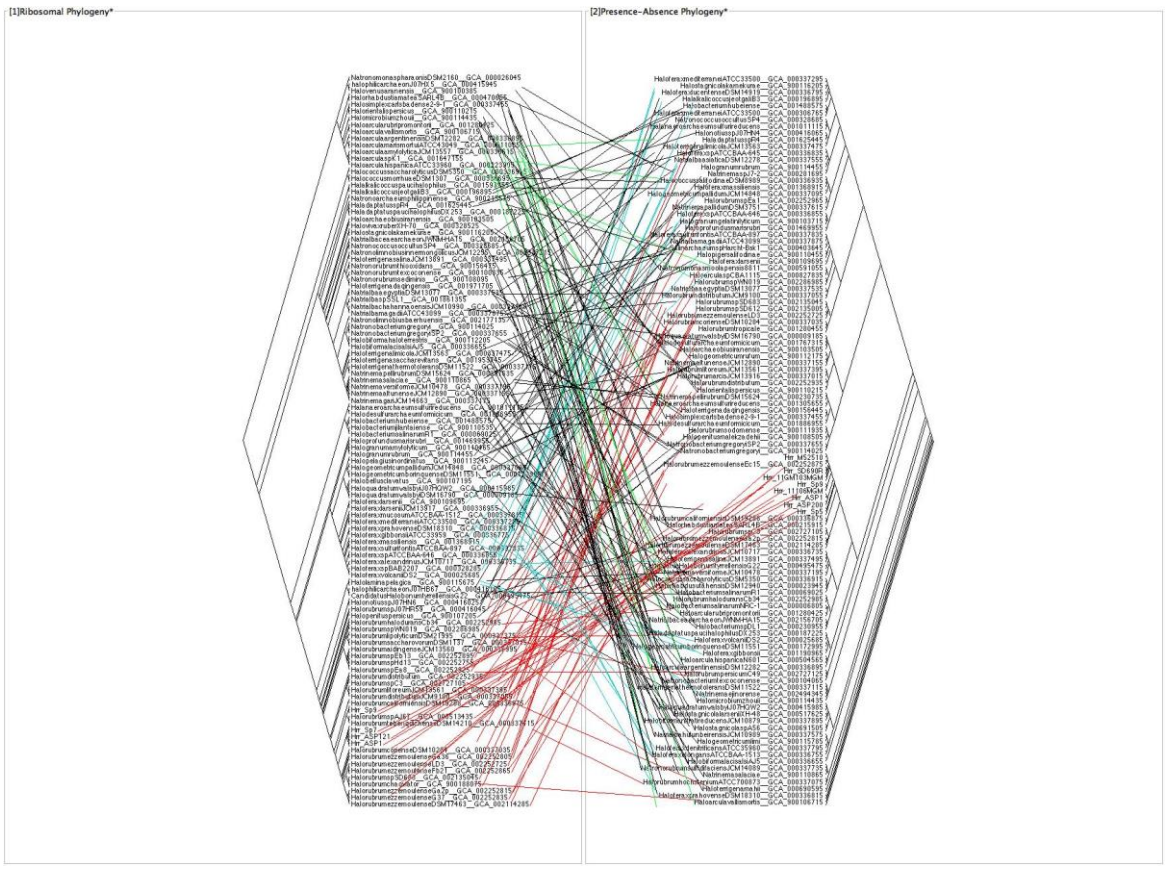


Figure08. Tanglegram matching the reference tree against the presence-absence tree. Some of the connecting lines for several taxonomic groups are colored to help illustrate the discord between the two topologies.

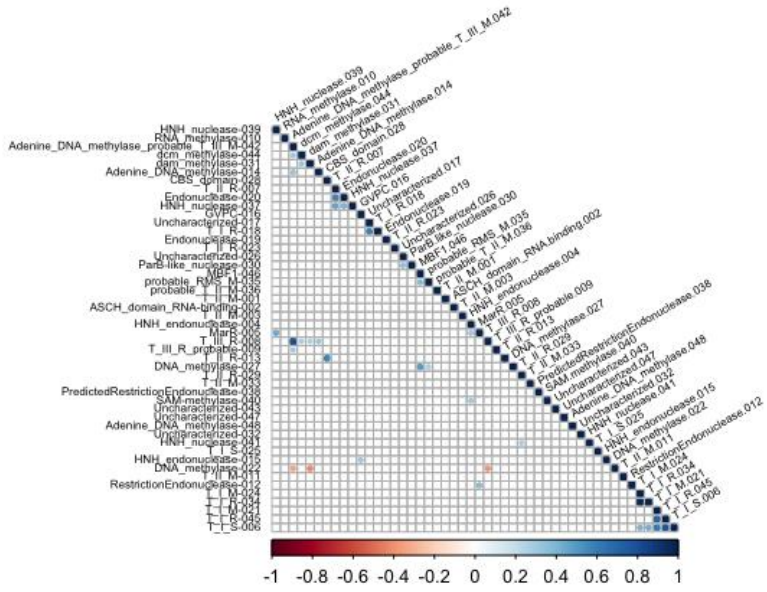


Figure09. Heatmap of the 48 RMS-candidate cHGs. Blue indicates significant positive correlation in the occurrence of the two cHGs. Red indicates a significant anti-correlation in the presence of the cHGs. Positive correlation indicates the cHGs co-occur while negative indicates that the presence of one means the other will not be present. Significance level is $p < 0.05$ with a Bonferroni correction applied for multiple tests.

Table01. Listing of genomes used in the study.

Name	NCBI ID	Origin
CandidatusHalobonumtyrrellensisG22	GCA_000495475	NCBI
Haladaptatuslitoreus	GCA_900156425	NCBI
HaladaptatuspaucihalophilusDX253	GCA_000187225	NCBI
HaladaptatuspaucihalophilusDX253	GCA_900142335	NCBI
HaladaptatusspR4	GCA_001625445	NCBI

HaladaptatusspW1	GCA_001723155	NCBI
HalalkalicoccusjeotgaliB3	GCA_000196895	NCBI
HalalkalicoccusjeotgaliB3	GCA_000337255	NCBI
Halalkalicoccuspaucihalophilus	GCA_001593955	NCBI
Halanaeroarchaeumsulfurireducens	GCA_001011115	NCBI
Halanaeroarchaeumsulfurireducens	GCA_001305655	NCBI
HalarchaeumacidiphilumMH1-52-1	GCA_000474235	NCBI
Haloarchaeobiusiranensis	GCA_900103505	NCBI
HaloarculaamylolyticaJCM13557	GCA_000336615	NCBI
HaloarculaargentinensisDSM12282	GCA_000336895	NCBI
HaloarculacaliforniaeATCC33799	GCA_000337755	NCBI
HaloarculahispanicaATCC33960	GCA_000223905	NCBI
HaloarculahispanicaN601	GCA_000504565	NCBI
HaloarculajaponicaDSM6131	GCA_000336635	NCBI
HaloarculamarismortuiATCC43049	GCA_000011085	NCBI
Haloarcularubripromontorii	GCA_001280425	NCBI
HaloarculasinaiiensisATCC33800	GCA_000337275	NCBI
HaloarculaspCBA1115	GCA_000827835	NCBI
HaloarculaspK1	GCA_001647155	NCBI
Haloarculavallismortis	GCA_900106715	NCBI
HaloarculavallismortisATCC29715	GCA_000337775	NCBI
Halobacteriumhubeiense	GCA_001488575	NCBI
Halobacteriumjilantaiense	GCA_900110535	NCBI
HalobacteriumsalinarumNRC-1	GCA_000006805	NCBI
HalobacteriumsalinarumR1	GCA_000069025	NCBI
HalobacteriumspDL1	GCA_000230955	NCBI
Halobaculumgomorrense	GCA_900129775	NCBI
Halobellusclavatus	GCA_900107195	NCBI
Halobiformahaloterrestris	GCA_900112205	NCBI
HalobiformalacisalsiAJ5	GCA_000226975	NCBI
HalobiformalacisalsiAJ5	GCA_000336655	NCBI
HalobiformanitratireducensJCM10879	GCA_000337895	NCBI
Halococcushamelinensis100A6	GCA_000336675	NCBI
HalococcusmorrhuaeDSM1307	GCA_000336695	NCBI
HalococcussaccharolyticusDSM5350	GCA_000336915	NCBI
HalococcussalifodinaeDSM8989	GCA_000336935	NCBI
HalococcusthailandensisJCM13552	GCA_000336715	NCBI
Halodesulfurarchaeumformicicum	GCA_001767315	NCBI
Halodesulfurarchaeumformicicum	GCA_001886955	NCBI
HaloferaxalexandrinusJCM10717	GCA_000336735	NCBI
HaloferaxdenitrificansATCC35960	GCA_000337795	NCBI

HaloferaxelongansATCCBAA-1513	GCA_000336755	NCBI
Haloferaxgibbonsii	GCA_001190965	NCBI
HaloferaxgibbonsiiATCC33959	GCA_000336775	NCBI
Haloferaxlarsenii	GCA_900109695	NCBI
HaloferaxlarseniiJCM13917	GCA_000336955	NCBI
HaloferaxlucentenseDSM14919	GCA_000336795	NCBI
Haloferaxmassiliensis	GCA_001368915	NCBI
HaloferaxmediterraneiATCC33500	GCA_000306765	NCBI
HaloferaxmediterraneiATCC33500	GCA_000337295	NCBI
HaloferaxmediterraneiATCC33500	GCA_000685635	NCBI
HaloferaxmucosumATCCBAA-1512	GCA_000337815	NCBI
HaloferaxprahovenseDSM18310	GCA_000336815	NCBI
HaloferaxspATCCBAA-644	GCA_000336975	NCBI
HaloferaxspATCCBAA-645	GCA_000336835	NCBI
HaloferaxspATCCBAA-646	GCA_000336855	NCBI
HaloferaxspBAB2207	GCA_000328285	NCBI
HaloferaxspSB3	GCA_001469875	NCBI
HaloferaxspSB29	GCA_001469865	NCBI
HaloferaxsulfurifontisATCCBAA-897	GCA_000337835	NCBI
HaloferaxvolcaniiDS2	GCA_000025685	NCBI
HaloferaxvolcaniiDS2	GCA_000337315	NCBI
HalogeometricumborinquenseDSM11551	GCA_000172995	NCBI
HalogeometricumborinquenseDSM11551	GCA_000337855	NCBI
Halogeometricumlimi	GCA_900115785	NCBI
HalogeometricumpallidumJCM14848	GCA_000337095	NCBI
Halogeometricumrufum	GCA_900112175	NCBI
Halogranumamylolyticum	GCA_900110465	NCBI
Halogranumgelatinilyticum	GCA_900103715	NCBI
Halogranumrubrum	GCA_900114455	NCBI
HalogranumsalariumB-1	GCA_000283335	NCBI
Halohastalitchfieldiae	GCA_900109065	NCBI
Halolaminapelagica	GCA_001307315	NCBI
Halolaminapelagica	GCA_900115675	NCBI
HalolaminaspCBA1230	GCA_002025255	NCBI
HalomicrobiummukohataeiDSM12286	GCA_000023965	NCBI
Halomicrobiumzhouii	GCA_900114435	NCBI
HalonotiusspJ07HN4	GCA_000416065	NCBI
HalonotiusspJ07HN6	GCA_000416025	NCBI
Halopelagiusinordinatus	GCA_900113245	NCBI
Halopelagiuslongus	GCA_900100875	NCBI
Halopenitusmalekzadehii	GCA_900108505	NCBI

Halopenituspersicus	GCA_900107205	NCBI
halophilicarchaeonJ07HB67	GCA_000416105	NCBI
halophilicarchaeonJ07HX5	GCA_000415945	NCBI
halophilicarchaeonJ07HX64	GCA_000416085	NCBI
Halopigersalifodinae	GCA_900110455	NCBI
HalopigerxanaduensisSH-6	GCA_000217715	NCBI
Haloplanusvescus	GCA_900107665	NCBI
Haloprofundusmarisrubri	GCA_001469955	NCBI
HaloquadratumspJ07HQX50	GCA_000416005	NCBI
HaloquadratumwalsbyiC23	GCA_000237865	NCBI
HaloquadratumwalsbyiDSM16790	GCA_000009185	NCBI
HaloquadratumwalsbyiJ07HQW1	GCA_000415965	NCBI
HaloquadratumwalsbyiJ07HQW2	GCA_000415985	NCBI
HalorhabdustiamateaSARL4B	GCA_000215915	NCBI
HalorhabdustiamateaSARL4B	GCA_000470655	NCBI
HalorhabdusutahensisDSM12940	GCA_000023945	NCBI
Halorientalispersicus	GCA_900110215	NCBI
Halorientalisregularis	GCA_900102305	NCBI
HalorientalisspIM1011	GCA_001989615	NCBI
HalorubrumaidingenseJCM13560	GCA_000336995	NCBI
HalorubrumarcisJCM13916	GCA_000337015	NCBI
HalorubrumcaliforniensisDSM19288	GCA_000336875	NCBI
Halorubrumchaoviator	GCA_900188075	NCBI
HalorubrumcorienseDSM10284	GCA_000337035	NCBI
Halorubrumdistributum	GCA_002252935	NCBI
HalorubrumdistributumJCM9100	GCA_000337055	NCBI
HalorubrumdistributumJCM10118	GCA_000337335	NCBI
HalorubrumezzemoulenseDSM17463	GCA_002114285	NCBI
HalorubrumezzemoulenseEc15	GCA_002252875	NCBI
HalorubrumezzemoulenseFb21	GCA_002252865	Papke Lab Sequencing
HalorubrumezzemoulenseG37	GCA_002252835	NCBI
HalorubrumezzemoulenseGa2p	GCA_002252815	NCBI
HalorubrumezzemoulenseGa36	GCA_002252805	NCBI
HalorubrumezzemoulenseLD3	GCA_002252725	NCBI
HalorubrumezzemoulenseLG1	GCA_002252735	NCBI
HalorubrumhaloduransCb34	GCA_002252985	NCBI
HalorubrumhochsteniumATCC700873	GCA_000337075	NCBI
HalorubrumkocuriiJCM14978	GCA_000337355	NCBI
HalorubrumlacusprofundiATCC49239	GCA_000022205	NCBI
HalorubrumlipolyticumDSM21995	GCA_000337375	NCBI
HalorubrumlitoreumJCM13561	GCA_000337395	NCBI

HalorubrumpersicumC49	GCA_002727125	NCBI
Halorubrumsaccharovorum	GCA_000700025	NCBI
HalorubrumsaccharovorumDSM1137	GCA_000337915	NCBI
Halorubrumsodomense	GCA_900111935	NCBI
HalorubrumspAJ67	GCA_000513435	NCBI
HalorubrumspC3	GCA_002727105	NCBI
HalorubrumspC191	GCA_002727095	NCBI
HalorubrumspE3	GCA_002252955	NCBI
HalorubrumspEa1	GCA_002252965	NCBI
HalorubrumspEa8	GCA_002252925	NCBI
HalorubrumspEb13	GCA_002252895	NCBI
HalorubrumspHd13	GCA_002252755	NCBI
Halorubrumsplb24	GCA_002252745	Papke Lab Sequencing
HalorubrumspJ07HR59	GCA_000416045	NCBI
HalorubrumspSD612	GCA_002135005	NCBI
HalorubrumspSD683	GCA_002135045	NCBI
HalorubrumspWN019	GCA_002286985	NCBI
HalorubrumtebenquichenseDSM14210	GCA_000337415	NCBI
HalorubrumterrestreJCM10247	GCA_000337435	NCBI
Halorubrumtropicale	GCA_001280455	NCBI
Halorubrumvacuolatum	GCA_900188065	NCBI
Halosimplexcarlsbadense2-9-1	GCA_000337455	NCBI
Halostagnicolakamekurae	GCA_900116205	NCBI
HalostagnicolalarseniiXH-48	GCA_000517625	NCBI
HalostagnicolaspA56	GCA_000691505	NCBI
Haloterrigenadaqingensis	GCA_001971705	NCBI
Haloterrigenadaqingensis	GCA_900156445	NCBI
Haloterrigenahispanica	GCA_900111485	NCBI
HaloterrigenalimicolaJCM13563	GCA_000337475	NCBI
Haloterrigenamahii	GCA_000690595	NCBI
Haloterrigenasaccharevitans	GCA_001953745	NCBI
HaloterrigenasalinaJCM13891	GCA_000337495	NCBI
HaloterrigenathermotoleransDSM11522	GCA_000337115	NCBI
HaloterrigenaturkmenicaDSM5511	GCA_000025325	NCBI
Halovenusaranensis	GCA_900100385	NCBI
HalovivaxasiaticusJCM14624	GCA_000337515	NCBI
HalovivaxruberXH-70	GCA_000328525	NCBI
Hrr_11GM103MGM		Papke Lab Sequencing
Hrr_11106MGM		Papke Lab Sequencing
Hrr_12103MGM		Papke Lab Sequencing
Hrr_ASP1		Papke Lab Sequencing

Hrr_ASP121		Papke Lab Sequencing
Hrr_ASP200		Papke Lab Sequencing
Hrr_M52510		Papke Lab Sequencing
Hrr_M425108A		Papke Lab Sequencing
Hrr_SD626R		Papke Lab Sequencing
Hrr_SD690R		Papke Lab Sequencing
Hrr_Sp3		Papke Lab Sequencing
Hrr_Sp5		Papke Lab Sequencing
Hrr_Sp7		Papke Lab Sequencing
Hrr_Sp9		Papke Lab Sequencing
NatrialbaaegyptiaDSM13077	GCA_000337535	NCBI
NatrialbaasiaticaDSM12278	GCA_000337555	NCBI
NatrialbaceaearchaeonJWNM-HA15	GCA_002156705	NCBI
NatrialbachahannaoensisJCM10990	GCA_000337135	NCBI
NatrialbahulunbeirensisJCM10989	GCA_000337575	NCBI
NatrialbamagadiiATCC43099	GCA_000025625	NCBI
NatrialbamagadiiATCC43099	GCA_000337875	NCBI
NatrialbaspSSL1	GCA_001861355	NCBI
NatrialbataiwanensisDSM12281	GCA_000337595	NCBI
NatrinemaaltunenseJCM12890	GCA_000337155	NCBI
Natrinemaejinorense	GCA_002494345	NCBI
NatrinemagariJCM14663	GCA_000337175	NCBI
NatrinemapallidumDSM3751	GCA_000337615	NCBI
NatrinemapellirubrumDSM15624	GCA_000230735	NCBI
NatrinemapellirubrumDSM15624	GCA_000337635	NCBI
Natrinemasalaciae	GCA_900110865	NCBI
NatrinemaspCBA1119	GCA_002572525	NCBI
NatrinemaspJ7-2	GCA_000281695	NCBI
NatrinemaversiformeJCM10478	GCA_000337195	NCBI
Natronoarchaeumphilippinense	GCA_900215575	NCBI
Natronobacteriumgregoryi	GCA_900114025	NCBI
NatronobacteriumgregoryiSP2	GCA_000230715	NCBI
NatronobacteriumgregoryiSP2	GCA_000337655	NCBI
Natronobacteriumtexcoconense	GCA_900104065	NCBI
NatronococcusamylolyticusDSM10524	GCA_000337675	NCBI
NatronococcusjeotgaliDSM18795	GCA_000337695	NCBI
NatronococcusoccultusSP4	GCA_000328685	NCBI
Natronolimnobiusbaerhuensis	GCA_002177135	NCBI
NatronolimnobiusinnermongolicusJCM12255	GCA_000337215	NCBI
Natronomonasmoolapensis8811	GCA_000591055	NCBI
NatronomonaspharaonisDSM2160	GCA_000026045	NCBI

NatronorubrumbangenseJCM10635	GCA_000337715	NCBI
Natronorubrumsediminis	GCA_900108095	NCBI
NatronorubrumsulfidifaciensJCM14089	GCA_000337735	NCBI
Natronorubrumtexcoconense	GCA_900100335	NCBI
Natronorubrumthiooxidans	GCA_900156475	NCBI
NatronorubrumtibetenseGA33	GCA_000337235	NCBI
SalinarchaeumspHarcht-Bsk1	GCA_000403645	NCBI

Table02. Listing of 48 RMS-candidate cHGs.

Homologous Group	arCOG Function
cHG_001	T_II_M-001
cHG_002	ASCH_domain_RNA-binding-002
cHG_003	T_II_M-003
cHG_004	HNH_endonuclease-004
cHG_005	MarR-005
cHG_006	T_I_S-006
cHG_007	T_II_R-007
cHG_008	T_III_R-008
cHG_009	T_III_R_probable-009
cHG_010	RNA_methylase-010
cHG_011	T_II_M-011
cHG_012	RestrictionEndonuclease-012
cHG_013	T_II_R-013
cHG_014	Adenine_DNA_methylase-014
cHG_015	HNH_endonuclease-015
cHG_016	GVPC-016
cHG_017	Uncharacterized-017
cHG_018	T_I_R-018
cHG_019	Endonuclease-019
cHG_020	Endonuclease-020
cHG_021	T_I_M-021
cHG_022	DNA_methylase-022
cHG_023	T_II_R-023
cHG_024	T_I_M-024
cHG_025	T_I_S-025
cHG_026	Uncharacterized-026

cHG_027	DNA_methylase-027
cHG_028	CBS_domain-028
cHG_029	T_II_R-029
cHG_030	ParB-like_nuclease-030
cHG_031	dam_methylase-031
cHG_032	Uncharacterized-032
cHG_033	T_II_M-033
cHG_034	T_I_R-034
cHG_035	probable_RMS_M-035
cHG_036	probable_T_II_M-036
cHG_037	HNH_nuclease-037
cHG_038	PredictedRestrictionEndonuclease-038
cHG_039	HNH_nuclease-039
cHG_040	SAM-methylase-040
cHG_041	HNH_nuclease-041
cHG_042	Adenine_DNA_methylase_probable_T_III_M-042
cHG_043	Uncharacterized-043
cHG_044	dcm_methylase-044
cHG_045	T_I_R-045
cHG_046	MBF1-046
cHG_047	Uncharacterized-047
cHG_048	Adenine_DNA_methylase-048

Table03. Important traits of cHGs with four or more ORFs. Column three lists the number of estimated horizontal transfer events. Columns five through 10 contain the top predicted recognition sites and the frequency of those predictions within the cHG.

cHG	#taxa	#transfers	Function	Recogsite #1	Frequency	Recogsite #2	Frequency	Recogsite #3	Frequency
cHG_001	16	9	T_II_M-001	GAAGGC	31%	GGRCA	31%		
cHG_003	38	21	T_II_M-003	CANCATC	53%	TAGGAG	21%		
cHG_004	12	4	HNH_endonuclease-004	GGCGCC	89%	GATC	11%		
cHG_006	61	44	T_I_S-006	GGAYNNNNNNTGG	24%	CAGNNNNNNTGCT	16%		
cHG_008	14	0	T_III_R-008	NA	100%				
cHG_010	55	15	RNA_methylase-010	ATTAAT	33%				
cHG_011	137	97	T_II_M-011	GCAAGG	49%	GKAAYG	28%		

cHG_012	8	5	RestrictionEndonuclease-012	GCGAA	29%	CAACNNNNNTC	29%	CTGGAG	29%
cHG_014	130	93	Adenine_DNA_methylase-014	GCAGG	45%	AAGCTT	32%		
cHG_015	21	13	HNH_endonuclease-015	GGCGCC	70%	YSCNS	15%		
cHG_016	12	6	GVPC-016	CANCATC	83%				
cHG_018	7	4	T_I_R-018	AACNNNNNNGTGC	73%	CTANNNNNNRTTC	27%		
cHG_019	4	3	Endonuclease-019	NA	100%				
cHG_021	88	58	T_I_M-021	GGAYNNNNNNTGG	37%	GTCANNNNNNRTCA	12%	CTCGAG	9%
cHG_022	290	120	DNA_methylase-022	CTAG	59%	CATTC	14%	CCCGGG	7%
cHG_023	37	28	T_II_R-023	NA	100%				
cHG_024	16	8	T_I_M-024	GAGNNNNNNVTGAC	75%	GACNNNNNNRTAC	19%		
cHG_025	4	2	T_I_S-025	GAGNNNNRTAA	75%	GAGNNNNNTAC	25%		
cHG_027	5	1	DNA_methylase-027	CATTC	100%				
cHG_030	4	2	ParB-like_nuclease-030	GATC	75%	CTAG	25%		
cHG_031	153	70	dam_methylase-031	GATC	70%	AB	22%		
cHG_032	116	60	Uncharacterized-032	GCAAGG	43%	GKAAYG	26%	GGTTAG	14%
cHG_033	66	38	T_II_M-033	CAARCA	40%	CTGAAG	36%		
cHG_034	16	11	T_I_R-034	GCANNNNNRTTA	69%	GGCANNNNNNTTC	19%		
cHG_035	19	9	probable_RMS_M-035	GGGAC	83%				
cHG_036	38	24	probable_T_II_M-036	CCWGG	42%	CCSGG	18%	GTAC	16%
cHG_037	6	4	HNH_nuclease-037						
cHG_039	5	4	HNH_nuclease-039	GGCGCC	100%				
cHG_041	6	4	HNH_nuclease-041						
cHG_042	21	8	Adenine_DNA_methylase_probable_T_III_M-042	RGTAAT	71%	NA	19%		
cHG_044	179	110	dcm_methylase-044	CGGCCG	24%	GTCGAC	13%	ACGT	11%
cHG_045	58	42	T_I_R-045	CCCNNNNNRTTGY	63%	GCANNNNNRTTA	28%		
cHG_048	54	35	Adenine_DNA_methylase-048	CCRGAG	36%	GTMKAC	30%		

Chapter 5 – A novel idea about how pan-genomes might evolve.

This chapter departs somewhat from the previously presented work and deals with the hypotheticals of evolution. The manuscript is an opinion piece postulating an extension of the well-known "Black Queen Hypothesis" (BQH) first posited by Morris et al. (Morris et al., 2012). Its crux is the notion that the pan-genome of a species, population, or community might act as a shared resource and each individual (or taxon) might not need to do everything on its own. The result might look like strong cooperation but the path to these ends could be reached through a race to cheat. This manuscript was prepared in collaboration with Shannon M. Soucy and J. Peter Gogarten. J. Peter Gogarten conceived of the original concept and wrote a grant application that became the basis for the 1st draft. Shannon M. Soucy participated in concept development and participated in the editing of the manuscript. I developed the concept with J. Peter Gogarten and developed the drafts from the original grant text, as well as editing of the manuscript. The major result of this chapter is the introduction of the "strong Black Queen Hypothesis" whereby mutual cheating might lead to apparent stable cooperation.

Chapter 5.1 The pan-genome as a shared genomic resource: mutual cheating, cooperation and the black queen hypothesis

frontiers in Microbiology

OPINION
published: 21 July 2015
doi: 10.3389/fmicb.2015.00728

The pan-genome as a shared genomic resource: mutual cheating, cooperation and the black queen hypothesis

Matthew S. Fullmer[1], *Shannon M. Soucy*[1] *and Johann Peter Gogarten*[1,2]*

[1] *Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA,* [2] *Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA*

Keywords: pan-genome, black queen hypothesis, red queen hypothesis, social cheating, gene transfer

OPEN ACCESS

Edited by:
Luis Delaye, Centro de Investigación y de Estudios Avanzados - Unidad Irapuato, Mexico

Reviewed by:
Luis David Alcaraz, Universidad Nacional Autónoma de México, Mexico
Luisa I. Falcon, Universidad Nacional Autónoma de México, Mexico

***Correspondence:**
Johann Peter Gogarten, gogarten@uconn.edu; jpgogarten@gmail.com

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *27 April 2015*
Accepted: *03 July 2015*
Published: *21 July 2015*

Citation:
Fullmer MS, Soucy SM and Gogarten JP (2015) The pan-genome as a shared genomic resource: mutual cheating, cooperation and the black queen hypothesis. Front. Microbiol. 6:728. doi: 10.3389/fmicb.2015.00728

Cells without Complete Genomes

Cells have long been recognized as life's building blocks (e.g., Virchow's dictum "*omnis cellula e cellula*," Virchow, 1860). Specifically, a cell's genome is considered the repository of genetic information that pairs with the cellular machinery to determine the organism's phenotype. Except for rare circumstances, the majority of a genome is passed on from ancestor to descendant, although the acquisition of genes from organisms that are not direct ancestors is recognized to play an important role in evolution (Swithers et al., 2012).

Jeffrey Lawrence, in discussing minimal genome size proposed a meta-cell model (Lawrence, 1999), in which many micelles (small vesicles containing resources, products, and genes) exchange genes frequently. Genes temporarily reside in a micelle and direct the synthesis of compounds important for replication. A micelle only can replicate when all compounds necessary for division have been generated. However, at each point in time only a fraction of the necessary genes are present in an individual micelle. This model relies on gene transfer being so frequent that each of the genes that encode necessary functions visits the individual micelles often enough to allow for sufficient synthesis of the necessary gene products for future micelle divisions. The meta-cell can be considered an organism, whose genome is divided into a network of micelles. Lawrence's meta-cell model is reminiscent of Woese's progenote (Woese, 1998) and Kandler's pre-cell populations (Kandler, 1994) that were postulated to have existed early in evolution before genes coalesced into genomes.

The Pan-Genome as a Shared Genomic Resource

For most bacterial and archaeal species different strains contain non-overlapping gene sets. The pan-genome of a taxon or group refers to the sum of all genes that are present in members of the group (Tettelin et al., 2005; Lapierre and Gogarten, 2009). Pan-genomes comprise the core genome, i.e., the genes that are found in all members, and the accessory genome, i.e., genes that are present in only one or a few members of the group. Welch et al. (2002) provided the first illustration that genome content in bacteria changes rapidly. Comparing three *Escherichia coli* strains they found the shared core to be less than 40% of the gene families present in all three genomes. More recently the size of this core was further reduced to only 6% of gene families present in 61 *E. coli* genomes (Lukjancenko et al., 2010). Baumdicker et al. (2012) estimate that the *Prochlorococcus* pan-genome contains about 58,000 genes, whereas the individual genomes encode only about 2000 genes each.

The pan-genome concept was originally developed to explore the fluidity of prokaryotic genomes (Tettelin et al., 2005). Because HGT is more frequent between close relatives (Andam and Gogarten, 2011), the pan-genome may also represent the set of genes that is potentially available via HGT to any member of the group. The function of the pan-genome may then be thought of as a shared resource. This is supported by the observation that genes encoding weakly selected functions are frequently lost from bacterial genomes, when they do not provide selective advantages, only to be re-acquired through HGT, when new conditions provide a selective advantage to carriers (Lawrence and Roth, 1996). The idea of the pan-genome of a population as a shared genomic resource is similar to the description of meta-cells and pre-cell populations. In particular, these concepts have in common that the individual genome of a cell or micelle does not represent a sufficient description of the genomic resources of the population. The following paragraphs discuss some factors that contribute to the large size of pan-genomes.

The Strong Black Queen Hypothesis

The black queen hypothesis proposed by Morris et al. (2012) is built on the premise of "leaky" common good functions, which cannot be restricted to benefit only the producer. The hypothesis suggests that these functions combined with selection for small genomes may lead to a situation in which these leaky functions are encoded in only a fraction of the genomes comprising the community. Under the black queen hypothesis a cell's evolution can follow one of two pathways (see **Figure 1**): (1) the cell can retain all genes encoding leaky functions (in the game of hearts, from which the name for the black queen hypothesis derives, this strategy is known as "shooting the moon"). The cost is a large genome maintaining and expressing many genes that are not essential to central metabolism, growth, and reproduction. Consequently, maintaining these genes and expression of extra proteins competes for cellular resources that could be put toward replication and results in a lower growth rate (Dong et al., 1995; Scott et al., 2010; Weiße et al., 2015). The advantage of the "shooting the moon" strategy is that following a population bottleneck all genes encoding leaky functions are available in the genome. These members of a community following this strategy may be thought of as analogous to a keystone species. (2) The cell looses some or all of its leaky functions and increases its growth rate (in hearts, this represents the usual strategy of taking as few point cards as possible). Traditionally this is described as cheating, as the second strategy relies upon other cells in the population for the leaky functions it has lost. If a bottleneck occurs, a single cheating cell is unlikely to survive on its own. A possible outcome of all cells in a population following strategy

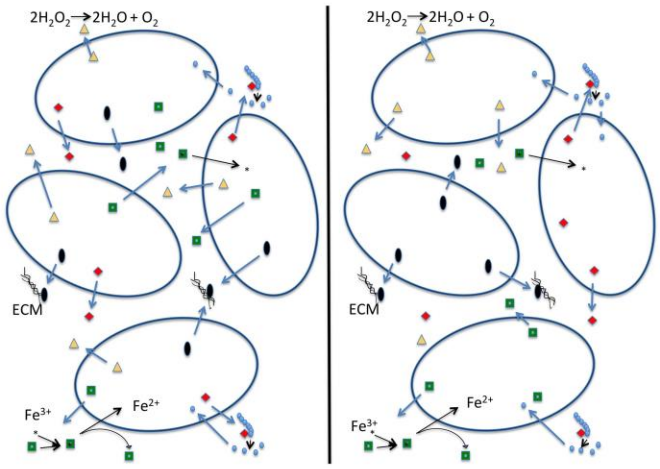


FIGURE 1 | Schematic depiction of a strong version of the black queen hypothesis (Morris et al., 2012). Each cell in the left panel contains all the genes that produce the four depicted common goods. Cells on the right, due to selection to minimize genome size, each produce only one of the common goods, and rely on the presence of other cells to produce the other ones. We consider this a strong version of the black queen hypothesis because it no longer contains exclusive helper strains; rather all individuals in the population described in the right hand panel are simultaneously helpers and beneficiaries. In this example dark open ovals represent individual cells; red diamonds: hydrolases that digest an extracellular polymer (e.g., phosphatase, sialidase, glucanase); green squares: siderophores that allow cells to acquire iron; black ovals: building blocks of the extracellular matrix (ECM) or enzymes that produce and assemble these building blocks; yellow triangles: enzymes that destroy oxygen radicals (e.g., catalase—reaction given; or peroxidase—reaction: ROOR' + electron donor (2 e^-) + $2H^+ \rightarrow$ ROH + R'OH).

#2 is that all members of a population cheat on some leaky functions. The members in the population then become mutually dependent on one another (**Figure 1**). In this scenario there are no keystone members providing all of the leaky functions. For the population to establish itself in a new environment several members of the population are required for the migration to be successful, as no single cell has all the components necessary to sustain itself. We term this the "strong" version of the black queen hypothesis. If all members of a population follow the second strategy, this may under some conditions lead to instability, the tragedy of the commons, and extinction of the population; however, experimental work by Morris et al. (2014) has shown that partitioning of a leaky common-goods function can enable the stable co-existence of two very similar organisms that use the same resources. Additionally, under natural conditions cells do rarely exist in homogeneous mixture (Davey and O'toole, 2000). Cells existing in biofilms or small aggregates are likely to be proximal to cells with which they share recent ancestry, and therefore proximal cells will have the same genotype with respect to shared functions. Drescher et al. (2014) show that *Vibrio cholera* can avoid the public goods dilemma by strengthening relationships between cells of the same genotype through creation of a thick biofilm, thereby providing a local selective advantage to producers of a particular common good in case this good becomes an overall limiting resource. It seems likely that genes encoding common goods are under frequency dependent selection, leading to local feedback loops that contribute to a long-term co-existence of the different types of cheaters.

Black vs. the Red Queen

Bacteria are under severe predation by phage (Thurber, 2009). They need to constantly change to evade predation, hence the analogy to the red queen from Lewis Carroll's (Carroll and Gardner, 1999) *Through the Looking-Glass*, who needs to run as fast as she can just to stay in place (Van Valen, 1973). The analysis of phage metagenomes and rank abundance curves indicated that the phage predation follows the *kill the winner* strategy (Hoffmann et al., 2007), where successful strains are targeted more frequently. The surprising long term stability of species composition despite phage predation suggests that cycling between different susceptible target cells occurs within a population and not between populations from different species (Rodriguez-Brito et al., 2010). Consequently, within a population, host genes that encode receptors utilized by phage and virus to enter the cell are expected to turn over quickly, creating within population diversity (Chaturongakul and Ounjai, 2014).

Random Acquisition of Genes

Genes are constantly acquired by genomes, and many of the transferred genes do not find a long term home in the recipient genome (Lawrence and Ochman, 1997). Among these genes are parasites (prophages) and selfish genetic elements. Most, but certainly not all (Lobkovsky et al., 2013), of the transferred genes are selectively neutral or nearly neutral to the recipient (Gogarten and Townsend, 2005; Baumdicker et al., 2010; Haegeman and Weitz, 2012). Though these genes may not find long term homes in the genomes they "visit," selfish genes especially can affect the rates of gene sharing and thus the size of the pan-genome in a population. Furthermore, many selfish elements induce genome rearrangements that can promote the loss and gain of genes, and thus may have a significant impact on the initiation of the loss of leaky functions.

Generation of paralogs may play a role in facilitation of loss of leaky functions. Additional copies increase gene dosage, ameliorating the loss of function in other members of the population by providing more of the common good. However, the pressure to delete genes from genomes is much stronger than to duplicate them (Mira et al., 2001) and an increase in gene transcription can have a similar or greater effect on the overall expression level (Weiße et al., 2015). Regardless of whether the increased production comes from paralogy or regulation it would need to be countered by a greater decrease in production from other common good functions to overcome the cost of increased protein expression.

Conclusion

Random acquisition of genes and selfish genetic elements, selection by predators, and cheating on common goods, all undoubtedly play a role in generating diversity within populations of bacteria and archaea. The conjecture of the strong black queen hypothesis is that mutual cheating leads to mutual dependencies and therefore cooperation. Under this hypothesis individual cells would be integrated into a meta-organism, whose genome is the pan-genome of the population, similar to Lawrence's meta-cells whose genome is distributed over individual micelles.

The pan-genome of a population as shared genomic resource could explain part of the "genome of Eden" paradox (Doolittle et al., 2003), where estimations of ancestral genomes are far larger and more complex than those of any extant individual genome. Large estimates of archaeal ancestors' genome sizes (Csurös and Miklós, 2009; Wolf et al., 2012) could actually represent the pan-genome of the ancestral population rather than any single cell. If this is the case, the complexity of the progenitor cells in a lineage/population might often be at a similar level of complexity as their extant relatives. We hypothesize the large estimates of progenitor genome size might in part reflect a "strong" black queen scenario where genome variation creates a large pan-genome, but no single cell contains a "keystone genome" with all genes in the population represented. More extensive studies of individual and population genomes, and rates of within population transfer are needed to confirm that master genomes, encoding all the leaky functions needed for survival of the population, can be and often are absent from a population.

If the hypothesis of the population pan-genome as a shared genomic resource is borne out, then the scientific community will need to continue to increase its appreciation for the import

of pan- and meta-genomes. Likewise, we may need to more seriously consider populations as the operative units in which genes are selected in rather than exclusively individual organisms. Similar to how Richard Dawkins (1976) advocated thinking of an organism as a collection of generally agreeable, but selfish, genes perhaps we should be thinking of lineages and populations as the collections of genes, i.e., pan-genomes, rather than the individual cells.

Acknowledgments

We thank Michael Stephens, Joerg Graf, and Thane Papke for Discussion. Work in the Gogarten-lab was supported through grants from the National Science Foundation (AToL DEB0830024), the National Aeronautics and Space Agency (Exobiology NNX13AI03G), and the Bi-national Science Foundation (BSF 2013061).

References

Andam, C. P., and Gogarten, J. P. (2011). Biased gene transfer in microbial evolution. *Nat. Rev. Microbiol.* 9, 543–555. doi: 10.1038/nrmicro2593

Baumdicker, F., Hess, W. R., and Pfaffelhuber, P. (2010). The diversity of a distributed genome in bacterial populations. *Ann. Appl. Probab.* 20, 1567–1606. doi: 10.1214/09-AAP657

Baumdicker, F., Hess, W. R., and Pfaffelhuber, P. (2012). The infinitely many genes model for the distributed genome of bacteria. *Genome Biol. Evol.* 4, 443–456. doi: 10.1093/gbe/evs016

Carroll, L., and Gardner, M. (1999). *The Annotated Alice: The Definitive Edition*. New York, NY: Norton.

Chaturongakul, S., and Ounjai, P. (2014). Phage-host interplay: examples from tailed phages and Gram-negative bacterial pathogens. *Front. Microbiol.* 5:442. doi: 10.3389/fmicb.2014.00442

Csurös, M., and Miklós, I. (2009). Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. *Mol. Biol. Evol.* 26, 2087–2095. doi: 10.1093/molbev/msp123

Davey, M. E., and O'toole, G. A. (2000). Microbial biofilms: from ecology to molecular genetics. *Microbiol. Mol. Biol. Rev.* 64, 847–867. doi: 10.1128/MMBR.64.4.847-867.2000

Dawkins, R. (1976). *The Selfish Gene*. Oxford: Oxford University Press.

Dong, H., Nilsson, L., and Kurland, C. G. (1995). Gratuitous overexpression of genes in *Escherichia coli* leads to growth inhibition and ribosome destruction. *J. Bacteriol.* 177, 1497–1504.

Doolittle, W. F., Boucher, Y., Nesbø, C. L., Douady, C. J., Andersson, J. O., and Roger, A. J. (2003). How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Philos. Trans. R. Soc. B Biol. Sci.* 358, 39–58. doi: 10.1098/rstb.2002.1185

Drescher, K., Nadell, C. D., Stone, H. A., Wingreen, N. S., and Bassler, B. L. (2014). Solutions to the public goods dilemma in bacterial biofilms. *Curr. Biol.* 24, 50–55. doi: 10.1016/j.cub.2013.10.030

Gogarten, J. P., and Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.* 3, 679–687. doi: 10.1038/nrmicro1204

Haegeman, B., and Weitz, J. S. (2012). A neutral theory of genome evolution and the frequency distribution of genes. *BMC Genomics* 13:196. doi: 10.1186/1471-2164-13-196

Hoffmann, K. H., Rodriguez-Brito, B., Breitbart, M., Bangor, D., Angly, F., Felts, B., et al. (2007). Power law rank-abundance models for marine phage communities. *FEMS Microbiol. Lett.* 273, 224–228. doi: 10.1111/j.1574-6968.2007.00790.x

Kandler, O. (1994). "The early diversification of life," in *Early Life on Earth*, ed S. Bengston (New York, NY: Columbia University Press), 152–509.

Lapierre, P., and Gogarten, J. P. (2009). Estimating the size of the bacterial pan-genome. *Trends Genet.* 25, 107–110. doi: 10.1016/j.tig.2008.12.004

Lawrence, J. G. (1999). "Gene transfer and minimal genome size," in *Size Limits of Very Small Microorganisms*, eds A. Knoll, M. J. Osborn, J. Baross, H. C. Berg, N. R. Pace, and M. Sogin (Washington, DC: National Research Council), 32–38.

Lawrence, J. G., and Ochman, H. (1997). Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* 44, 383–397. doi: 10.1007/PL00006158

Lawrence, J. G., and Roth, J. R. (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* 143, 1843–1860.

Lobkovsky, A. E., Wolf, Y. I., and Koonin, E. V. (2013). Gene frequency distributions reject a neutral model of genome evolution. *Genome Biol. Evol.* 5, 233–242. doi: 10.1093/gbe/evt002

Lukjancenko, O., Wassenaar, T. M., and Ussery, D. W. (2010). Comparison of 61 sequenced *Escherichia coli* genomes. *Microb. Ecol.* 60, 708–720. doi: 10.1007/s00248-010-9717-3

Mira, A., Ochman, H., and Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends Genet.* 17, 589–596. doi: 10.1016/S0168-9525(01)02447-7

Morris, J. J., Lenski, R. E., and Zinser, E. R. (2012). The Black Queen hypothesis: evolution of dependencies through adaptive gene loss. *MBio* 3:e00036-12. doi: 10.1128/mbio.00036-12

Morris, J. J., Papoulis, S. E., and Lenski, R. E. (2014). Coexistence of evolving bacteria stabilized by a shared black queen function. *Evolution* 68, 2960–2971. doi: 10.1111/evo.12485

Rodriguez-Brito, B., Li, L., Wegley, L., Furlan, M., Angly, F., Breitbart, M., et al. (2010). Viral and microbial community dynamics in four aquatic environments. *ISME J.* 4, 739–751. doi: 10.1038/ismej.2010.1

Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z., and Hwa, T. (2010). Interdependence of cell growth and gene expression: origins and consequences. *Science* 330, 1099–1102. doi: 10.1126/science.1192588

Swithers, K. S., Soucy, S. M., and Gogarten, J. P. (2012). The role of reticulate evolution in creating innovation and complexity. *Int. J. Evol. Biol.* 2012:418964. doi: 10.1155/2012/418964

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U.S.A.* 102, 13950–13955. doi: 10.1073/pnas.0506758102

Thurber, R. V. (2009). Current insights into phage biodiversity and biogeography. *Curr. Opin. Microbiol.* 12, 582–587. doi: 10.1016/j.mib.2009.08.008

Van Valen, L. (1973). A new evolutionary law. *Evol. Theory* 1, 1–30.

Virchow, R. L. K. (1860). *Cellular Pathology (Google eBook)*. London: John Churchill. Available online at: https://play.google.com/store/books/details/Rudolf_Ludwig_Karl_Virchow_Cellular_Pathology?id=JUth7ntb0_AC

Weiße, A. Y., Oyarzún, D. A., Danos, V., and Swain, P. S. (2015). Mechanistic links between cellular trade-offs, gene expression, and growth. *Proc. Natl. Acad. Sci. U.S.A.* 112, E1038–E1047. doi: 10.1073/pnas.1416533112

Welch, R. A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., et al. (2002). Extensive mosaic structure revealed by the

complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* 99, 17020–17024. doi: 10.1073/pnas.2525 29799

Woese, C. (1998). The universal ancestor. *Proc. Natl. Acad. Sci. U.S.A.* 95, 6854–6859. doi: 10.1073/pnas.95. 12.6854

Wolf, Y. I., Makarova, K. S., Yutin, N., and Koonin, E. V. (2012). Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. *Biol. Direct* 7:46. doi: 10.1186/1745-61 50-7-46

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Chapter 6 - Conclusions

Identifying and delimiting taxonomic units, usually at the species level, has been a prevalent theme of this dissertation. Chapter 2 features two first-author publications centered heavily around this topic (Colston et al., 2014; Fullmer et al., 2014b). Likewise, in three publications in the appendices my contributions centered around delimiting and classifying taxa (Collins et al., 2015; Gromek et al., 2016; Ram Mohan et al., 2014). My work in these manuscripts emphasizes established methods in the field (Auch et al., 2010; Konstantinidis and Tiedje, 2005; Papke et al., 2007; Sullivan et al., 2005) and utilizes them in largely orthodox manners. These works have some technical novelty by being an early use of average nucleotide identity (ANI) in the Halobacteria, as well as introducing a new multilocus sequence analysis scheme and recommendations for utilizing whole genome comparisons in the *Aeromonas*. Chapter 3 covers my work developing a novel method to extend existing ANI methodology. Chapter 3 also continues the classification theme with work demonstrating a method with the potential to delimit taxonomic ranks *in silico* above species.

The fourth chapter reports a deeper examination of the Halobacteria. The first section is a book chapter wherein I discuss the prevalence, role, and evolutionary impact of horizontal gene transfer in the Halobacteria (Fullmer et al., 2014a). The second section reports my development and implementation of a search

methodology for identifying restriction-methylation genes in the Halobacteria. After identifying them, I present evidence demonstrating their frequent horizontal transfer. These results set the stage for possible future work examining their impact on divergence in populations of *Halorubrum*.

The fifth chapter covers a different type of work than the proceeding chapter. I present a hypothesis extending the Black Queen Hypothesis (Morris et al., 2012). This hypothesis, termed the "strong" Black Queen, proposes how mutual cheating might result in a stable community with mixed production of common goods (Fullmer et al., 2015). Perhaps the most interesting consequence of this hypothesis is how pure cheating can create a final result possibly indistinguishable from mutualistic relationships.

These chapters are united by only one theme. That is a desire to understand how microbes evolve. The motivation for exploring the classification of microbes lies in needing to understand where evolution has placed them before the question of how they arrived there can be asked. Chapter 4 is prominently features examination of forces that are known, or are proposed, to shape the evolutionary history of the Halobacteria. Parts of chapter 1 (Fullmer et al., 2014b) and the appendices (Ram Mohan et al., 2014; Soucy et al., 2014) also explore the forces that shape the group. Chapter 5 directly ponders what forces could be shaping real populations and communities and offers an idea for further research to test.

References

Auch, A. F., Jan, M. von, Klenk, H.-P., and Göker, M. (2010). Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand. Genomic Sci.* 2, 117–134. doi:10.4056/sigs.531120.

Collins, A. J., Fullmer, M. S., Gogarten, J. P., and Nyholm, S. V. (2015). Comparative genomics of Roseobacter clade bacteria isolated from the accessory nidamental gland of Euprymna scolopes. *Front. Microbiol.* 6. doi:10.3389/fmicb.2015.00123.

Colston, S. M., Fullmer, M. S., Beka, L., Lamy, B., Gogarten, J. P., and Graf, J. (2014). Bioinformatic Genome Comparisons for Taxonomic and Phylogenetic Assignments Using Aeromonas as a Test Case. *mBio* 5, e02136-14. doi:10.1128/mBio.02136-14.

Fullmer, M. S., Gogarten, J. P., and Papke, R. T. (2014a). "Horizontal Gene Transfer in Halobacteria," in *Halophiles: Genetics and Genomes* (Caister Academic Press), 196.

Fullmer, M. S., Soucy, S. M., and Gogarten, J. P. (2015). The pan-genome as a shared genomic resource: mutual cheating, cooperation and the black queen hypothesis. *Evol. Genomic Microbiol.*, 728. doi:10.3389/fmicb.2015.00728.

Fullmer, M. S., Soucy, S. M., Swithers, K. S., Makkay, A. M., Wheeler, R., Ventosa, A., et al. (2014b). Population and genomic analysis of the genus Halorubrum. *Extreme Microbiol.* 5, 140. doi:10.3389/fmicb.2014.00140.

Gromek, S. M., Suria, A. M., Fullmer, M. S., Garcia, J. L., Gogarten, J. P., Nyholm, S. V., et al. (2016). Leisingera sp. JC1, a Bacterial Isolate from Hawaiian Bobtail Squid Eggs, Produces Indigoidine and Differentially Inhibits Vibrios. *Front. Microbiol.* 7. doi:10.3389/fmicb.2016.01342.

Konstantinidis, K. T., and Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 102, 2567–2572. doi:10.1073/pnas.0409727102.

Morris, J. J., Lenski, R. E., and Zinser, E. R. (2012). The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. *mBio* 3, e00036-12. doi:10.1128/mBio.00036-12.

Papke, R. T., Zhaxybayeva, O., Feil, E. J., Sommerfeld, K., Muise, D., and Doolittle, W. F. (2007). Searching for species in haloarchaea. *Proc. Natl. Acad. Sci.* 104, 14092–14097. doi:10.1073/pnas.0706358104.

Ram Mohan, N., Fullmer, M. S., Makkay, A. M., Wheeler, R. W., Ventosa, A., Naor, A., et al. (2014). Evidence from phylogenetic and genome fingerprinting analyses suggests rapidly changing variation in Halorubrum and Haloarcula populations. *Extreme Microbiol.* 5, 143. doi:10.3389/fmicb.2014.00143.

Soucy, S. M., Fullmer, M. S., Papke, R. T., and Gogarten, J. P. (2014). Inteins as indicators of gene flow in the halobacteria. *Extreme Microbiol.* 5, 299. doi:10.3389/fmicb.2014.00299.

Sullivan, C. B., Diggle, M. A., and Clarke, S. C. (2005). Multilocus sequence typing. *Mol. Biotechnol.* 29, 245–254. doi:10.1385/MB:29:3:245.

Appendices – Non 1st-author peer-reviewed publications

Appendix A – Ram Mohan et al., 2014

Evidence from phylogenetic and genome fingerprinting analyses suggests rapidly changing variation in *Halobrubrum* and *Haloarcula* populations

This section features Nikhil Ram Mohan's manuscript on using RAPD to quickly identify genomic heterogeneity in population isolates (Ram Mohan et al., 2014). R. Thane Papke, J. Peter Gogarten, and Antonio Ventosa conceived the research. Nikhil Ram Mohan, Matthew S. Fullmer, Andrea M. Makkay, and Ryan Wheeler gathered data, and performed the analyses. Nikhil Ram Mohan, Matthew S. Fullmer, Andrea M. Makkay, Ryan Wheeler, Antonio Ventosa, J. Peter Gogarten, and R. Thane Papke wrote the manuscript.

frontiers in MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 09 April 2014
doi: 10.3389/fmicb.2014.00143

Evidence from phylogenetic and genome fingerprinting analyses suggests rapidly changing variation in *Halorubrum* and *Haloarcula* populations

Nikhil Ram Mohan[1], Matthew S. Fullmer[1], Andrea M. Makkay[1], Ryan Wheeler[1], Antonio Ventosa[2], Adit Naor[3], J. Peter Gogarten[1] and R. Thane Papke[1]*

[1] *Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA*
[2] *Department of Microbiology and Parasitology, University of Seville, Seville, Spain*
[3] *Molecular Microbiology and Biotechnology, Tel Aviv University, Tel Aviv, Israel*

Edited by:
Jesse Dillon, California State University, Long Beach, USA

Reviewed by:
Jocelyne DiRuggiero, The Johns Hopkins University, USA
James A. Coker, University of Maryland, University College, USA

Correspondence:
R. Thane Papke, Department of Molecular and Cell Biology, University of Connecticut, 91 N. Eagleville Rd., Storrs, CT 06269, USA
e-mail: thane@uconn.edu

Halobacteria require high NaCl concentrations for growth and are the dominant inhabitants of hypersaline environments above 15% NaCl. They are well-documented to be highly recombinogenic, both in frequency and in the range of exchange partners. In this study, we examine the genetic and genomic variation of cultured, naturally co-occurring environmental populations of Halobacteria. Sequence data from multiple loci (~2500 bp) identified many closely and more distantly related strains belonging to the genera *Halorubrum* and *Haloarcula*. Genome fingerprinting using a random priming PCR amplification method to analyze these isolates revealed diverse banding patterns across each of the genera and surprisingly even for isolates that are identical at the nucleotide level for five protein coding sequenced loci. This variance in genome structure even between identical multilocus sequence analysis (MLSA) haplotypes indicates that accumulation of genomic variation is rapid: faster than the rate of third codon substitutions.

Keywords: Halobacteria, MLSA, genome fingerprinting, Aran-Bidgol lake, environmental population

INTRODUCTION

Members of the class Halobacteria (Domain: Archaea; Phylum: Euryarchaeota) are the dominant inhabitants of hypersaline environments (Anton et al., 1999; Ghai et al., 2011). These hypersaline environments provide extreme growth conditions in the form of high salinity and ionic concentrations with variations in pH, and temperature (Oren, 2002). Such extreme conditions are necessary for Halobacteria, also called haloarchaea, to live. The environment is also subject to low solubility of gases, low diffusion rates, and very low water activity (Litchfield, 1998). To overcome many of these obstacles, haloarchaea can generate ATP from light energy (Lozier et al., 1975) and have gas vesicles to buoyantly lift themselves to the surface (Jones et al., 1991). Osmotic survival in these brines is managed by maintaining a cytosolic salinity in equilibrium with that of the environment, a feat that requires solubilized proteins under those conditions, and solved with a proteome enriched in acidic and depleted of basic amino acids (Oren, 2002).

Haloarchaea have a well-documented capacity for generating enormous amounts of genetic variation through horizontal gene transfer (HGT) (Papke et al., 2004, 2007; Cuadros-Orellana et al., 2007; Lynch et al., 2012; Naor et al., 2012; Williams et al., 2012; Demaere et al., 2013; Podell et al., 2013). From the very first genome sequence analysis of *Halobacterium* strain NRC-1, evidence was provided for the acquisition of aerobic respiration genes via HGT from Bacteria (Ng et al., 2000). Since then, several studies on specific genes of interest [e.g., rhodopsins (Sharma et al., 2007), ribosomal RNAs (Boucher et al., 2004), and tRNA synthetases (Andam et al., 2012)] have further demonstrated gene transfer into and among the haloarchaea. A recent report suggested that this process of generating diversity has been ongoing since before the group's last universal common ancestor and that HGT played a huge role in changing their physiology from an autotrophic anaerobe to a heterotrophic aerobe (Nelson-Sathi et al., 2012). Population genetics analysis on strains from the genus *Halorubrum* using multilocus sequence analysis (MLSA) demonstrated that alleles at different loci are unlinked indicating that homologous recombination (HR) is frequent enough within phylogenetically defined groups to randomize traits among individuals (Papke et al., 2004, 2007), an observation once considered unique to sexually reproducing eukaryotes. Analysis of 20 haloarchaeal genomes showed that there are no absolute barriers to HR, which occurs regularly and proportionally to genetic distance throughout the haloarchaea (Williams et al., 2012). Community analyses using metagenomics revealed that genes are coming and going quickly within *Haloquadratum walsbyi* populations, suggesting there may be very few identical genomes within the species (Legault et al., 2006; Cuadros-Orellana et al., 2007). Perhaps most striking is their ability to exchange large swaths of genetic information. Mating experiments between *Haloferax volcanii* and *Haloferax mediterranei* demonstrated between ~10 and 18% (~300–500 kb) of their chromosome could be transferred in a single fragment (Naor et al., 2012). Also, genomes of highly divergent strains (e.g., <75% average nucleotide identity) isolated from Deep Lake, Antarctica were shown to share many ~100% identical DNA

sequences in fragments up to 35 Kb in length (Demaere et al., 2013).

MLSA has often been used as a technique for classifying microorganisms (Maiden et al., 1998), including halophiles (Papke et al., 2011; De la Haba et al., 2012), but it is also used to estimate population variation and gene flow (Feil et al., 2000). Assumptions using MLSA regarding how representative multiple genes are for capturing individual variation, and thus the appearance of clonality, can lead to erroneous conclusions. For instance, two strains may have identical sequences across multiple loci, but unexamined genomic variation might be high and belie the interpretation of little or no recombination. Indeed, studies are demonstrating that there are vast amounts of variation within bacterial species/populations. Environmental isolates with identical HSP-60 genes from a natural coastal *Vibrio* sp. population demonstrated that the overwhelming majority of individual strains were unique as determined by chromosome pulse field gel electrophoresis, with some strains differing by up to a megabase in genome size (Thompson et al., 2005). This variation in genome size and the existence of "open" (i.e., infinite) pan-genomes like that of *Prochlorococcus marinus* and others (Tettelin et al., 2008; Lapierre and Gogarten, 2009) suggest that HGT is so frequent that for at least some species every cell may be genetically distinct.

To get a better understanding for the genomic variation within closely related haloarchaeal strains we examined naturally co-occurring environmental strains from the genera *Halorubrum* and *Haloarcula* isolated from the Aran-Bidgol salt lake in Iran. We used MLSA to identify closely related strains, and a PCR genome fingerprinting technique that randomly primed amplification sites along the chromosome to generate a gel electrophoresis pattern that enabled us to inexpensively compare genomic variation of the isolates.

MATERIALS AND METHODS

GROWTH CONDITIONS AND DNA EXTRACTION

Aran-Bidgol *Halorubrum* and *Haloarcula* spp. cultures were grown in Hv-YPC medium (Allers et al., 2004) at 37°C with agitation. DNA from haloarchaea was isolated as described in the Halohandbook (http://www.haloarchaea.com/resources/halohandbook/). Briefly, stationary-phase cells were pelleted at 10,000 ×*g*, supernatant was removed and the cells were lysed in distilled water. An equal volume of phenol was added, and the mixture was incubated at 65°C for 1 h prior to centrifugation to separate the phases. The aqueous phase was reserved and phenol extraction was repeated without incubation, and followed with a phenol/chloroform/iso-amyl alcohol (25:24:1) extraction. The DNA was precipitated with ethanol, washed, and resuspended in TE (10 mM tris, pH 8.0, 1 mM EDTA). Type strains were grown, and DNA was purified as described by Papke et al. (2011).

SEQUENCE ACQUISITION FOR MLSA

Five housekeeping genes were amplified using PCR. The loci were *atpB*, *ef-2*, *glnA*, *ppsA*, and *rpoB* and the primers used for each locus are listed in **Table 1**. To more efficiently sequence PCR products, an 18 bp M13 sequencing primer was added to the 5′ end of each degenerate primer (**Table 1**). Each PCR reaction was 20 μl in volume. Phire Hot Start II DNA polymerase (Thermo Scientific) was used in the amplification reactions. The PCR reaction was run on a Mastercycler Ep Thermocycler (Eppendorf) using the following PCR cycle protocol: 30 s initial denaturation at 98°C, followed by 40 cycles of 30 s at 98°C, 5 s at the annealing temperature for each set of primers, and 15 s at 72°C. Final elongation occurred at 72°C for 1 min. **Table 2** provides a detailed list of reagents and the PCR mixtures for each amplified locus. The PCR products were separated by gel electrophoresis with agarose (1%). Gels were stained with ethidium bromide. An exACTGene mid-range plus DNA ladder (Fisher Scientific International Inc.) was used to estimate the size of the amplicons, which were purified using Wizard SV gel and PCR cleanup system (Promega). The purified amplicons were sequenced by Genewiz Inc. The sequences obtained for the five genes in this study were submitted to Genbank under the following accession numbers: KJ152221–KJ152260, KJ152261–KJ152298, KJ152362–KJ152397, KJ152398–KJ152433, and KJ152323–KJ152361.

PHYLOGENETIC ANALYSIS

Type strain genomes were obtained from the NCBI ftp repository. Blast searches identified DNA top hits for each MLSA target gene (*atpB*, *ef-2*, *glnA*, *ppsA*, and *rpoB*) in each genome. Multiple-sequence alignments (MSAs) were created from the DNA genome hits as well as the PCR amplicons using MUSCLE (Edgar, 2004) (alignments available upon request) with its refine function. The

Table 1 | Degenerate primers used to PCR amplify and sequence the *atpB, ef-2, glnA, ppsA,* and *rpoB* genes for MLSA.

	MLSA primer sequence 5′–3′	
Locus	**Forward**	**Reverse**
atpB	tgt aaa acg acg gcc agt aac ggt gag scv ats aac cc	cag gaa aca gct atg act tca ggt cvg trt aca tgt a
ef-2	tgt aaa acg acg gcc agt atc cgc gct bta yaa stg g	cag gaa aca gct atg act ggt cga tgg wyt cga ahg g
glnA	tgt aaa acg acg gcc agt cag gta cgg gtt aca sga cgg	cag gaa aca gct atg acc ctc gcs ccg aar gac ctc gc
ppsA	tgt aaa acg acg gcc agt ccg cgg tar ccv agc atc gg	cag gaa aca gct atg aca tcg tca ccg acg arg gyg g
rpoB	tgt aaa acg acg gcc agt tcg aag agc cgg acg aca tgg	cag gaa aca gct atg acc ggt cag cac ctg bac cgg ncc

Table 2 | PCR conditions for each locus.

	atpB	ef-2	glnA	ppsA	rpoB
Water (μl)	11.6	8.2	11.8	7.9	11.9
5× phire reaction buffer (μl)	4.0	4.0	4.0	4.0	4.0
DMSO (μl)	0.6	0	0.4	0.6	0.6
Acetamide (25%)	0	4.0	0	4.0	0
dNTP mix (10 mM)	0.4	0.4	0.4	0.4	0.4
Forward primer (10 mM)	1.0	1.0	1.0	1.0	1.0
Reverse primer (10 mM)	1.0	1.0	1.0	1.0	1.0
Phire hot start II DNA polymerase (μl)	0.4	0.4	0.4	0.4	0.4
Template DNA (20 ng μl^{-1})	1.0	1.0	1.0	0.7	0.7
Annealing temperature (°C)	60.0	61.0	69.6	66.0	63.7

MSA length was manually trimmed down to the lengths of the PCR amplicons. In-house scripts created a concatenated alignment of all five genes. A model of evolution was determined using the Akaike Information Criterion with correction for small sample size (AICc). The jModelTest 2.1.4 (Darriba et al., 2012) program was used to compute likelihoods from the nucleotide alignment and to perform the AICc test (Akaike, 1974). The AICc reported the best-fitting model to be GTR + Gamma estimation + Invariable site estimation. A maximum likelihood (ML) phylogeny was generated from the concatenated MSA using the PhyML v3.0_360–500 (Guindon et al., 2010). The model used in PhyML corresponded to the one favored by jModeltest: GTR model, estimated p-invar, 4 substitution rate categories, estimated gamma distribution with 100 bootstrap replicates. The number of nucleotide differences in pairwise comparisons were determined using MEGA 5 (Tamura et al., 2011).

GENOMIC FINGERPRINTING

In total, DNA from 81 haloarchaeal type strains and 43 isolates from the Aran-Bidgol lake were tested. Each primer selected has successfully been used in genome fingerprinting in previous studies. Primers P1 and P2 were used to fingerprint *Vibrio harveyi* bacteriophages (Shivu et al., 2007), primers OPA-9 and OPA-13 were used to asses marine viral richness (Winget and Wommack, 2008). The last primer, FALL-A was adapted from the primer used (Barrangou et al., 2002; Winget and Wommack, 2008) to study bacteriophages isolated from an industrial sauerkraut fermentation. Amplification conditions for each strain were equal to enable accurate comparison between banding patterns obtained. Each sample was diluted to 20 ng μl^{-1} and amplified within the following reaction mixture: 12.5 μl SYBR Universal Faststart Mastermix (Roche), 4.5 μl dH_20, 1.5 μl for each of five primers at 10 ng μl^{-1} (see **Table 3**), and 0.5 μl of template DNA. Two thermocycler programs were used in succession. The first included an initial 10 min denaturation at 94°C, followed by 4 cycles of a 45 s denaturation also at 94°C, annealing at 30°C for 2 min, and extension at 72°C for 50 s. This was followed by another 35 cycle program: 94°C for 17 s, 36°C for 30 s, and 72°C for 45 s, and a final extension for 10 min at 72°C. The aim of these repeated programs with low annealing temperatures and long annealing times is to produce as many non-specific bands as possible for each sample, increasing the resolving power of the method. Strains were amplified in triplicate to ensure that a repeatable banding pattern could be obtained.

Table 3 | Random primers for genomic fingerprinting.

Primers	
Primer name	**Sequence**
P1	5′-CCGCAGCCAA-3′
P2	5′-ACGGGCAGC-3′
OPA-9	5′-GGGTAACGCC-3′
OPA-13	5′-CAGCAGCCAC-3′
FALL-A	5′-ACGCGCCCTG-3′

GEL ELECTROPHORESIS

Reactions mixtures from PCR experiments were held at 4°C prior to electrophoresis. Standard DNA electrophoresis was carried out with replicates from each strain. Gels were 1.5% agarose and run at 12 v for 16 h at 4°C with the goal of producing crisp bands easily distinguishable by the analysis software. Gels were stained with ethidium bromide prior to imaging.

IMAGING AND ANALYSIS

A digital image of each gel was created using a GelDoc (UVP). Images were then analyzed using the Phoretix 1D Pro program from the TotalLab Inc. (www.totallab.com). Banding patterns were standardized for cross gel comparisons by calibrating Rf lines on individual gels. Phoretix 1D Pro converts banding patterns into a format that can be used to produce a dendrogram comparing the differences and similarities between the patterns of amplicons. The final dendrogram was created within Phoretix 1D Pro using UPGMA statistical analysis on Dice coefficients (Dice, 1945) for each of the lanes. A measure of the correlation between the matrix similarities and the dendrogram derived similarities, the cophenetic correlation coefficients (Sokal and Rohlf, 1962) were determined for each sub-cluster of the dendrogram and displayed on the nodes of the constructed dendrograms to estimate the robustness of each cluster.

RESULTS

GENOMIC FINGERPRINTING

The repeatability of banding patterns, and thus the success of the fingerprinting technique was tested on 81 haloarchaeal type strains. The PCR on each of the 81 strains was run in triplicate and the products were run on adjacent wells. **Figure 1** demonstrates results of the banding pattern for 18 out of the 81 type strains, 15 from the genus *Halorubrum,* and one each from the genus *Halosarcina, Halosimplex,* and *Halostagnicola.* Repeatability for the other 63 was examined and they were consistent, as in **Figure 1** (data are not shown). Repeatability of the technique indicated robustness of the conditions and primers used and provide confidence for estimating variation between strains.

We were interested to know if the random primers can be used as a screening technique. If banding patterns could reliably demonstrate similarity within genera for instance, newly cultured yet unidentified strains could be easily screened and a general taxonomic decision could be made. Therefore, the banding patterns for the 81 total haloarchaeal type strains were assessed using software that produced a dendrogram of the genomic fingerprints. **Figure 2** is the UPGMA dendrogram determined for the above type strains. Compared to other studies (e.g., Shivu et al., 2007; Winget and Wommack, 2008), our genome fingerprinting technique offers very little banding pattern complexity. There are two possible reasons—the primers were designed for systems other than the haloarchaea and adopted for our purposes, and PCR bias, though if it occurs is reproducible (see **Figure 1**). Yet, species specific banding patterns observed earlier in haloarchaea (Martinez-Murcia and Rodriguez-Valera, 1994) are also observed here; each species appears to have a unique banding pattern. However, there is very little clustering at the genus level. For instance, some species within the same genus

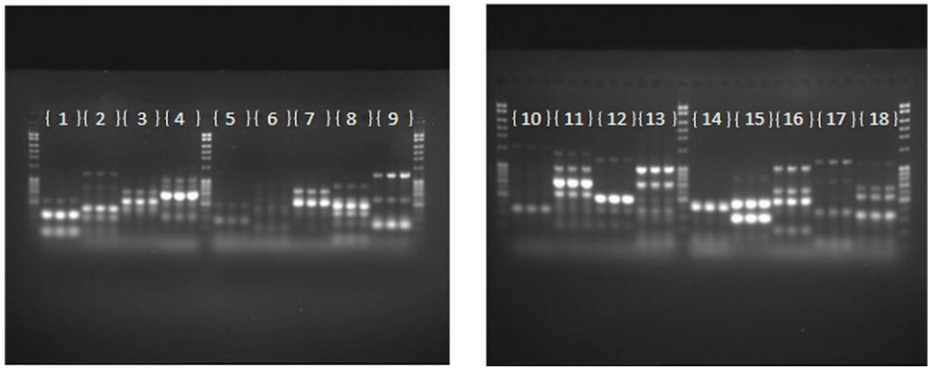


FIGURE 1 | Repeatability of the fingerprinting technique. Each number represents a type strain analyzed in triplicate. (1) *Halorubrum arcis* JCM 13916 (2) *Halorubrum coriense* DSM 10284 (3) *Halorubrum distributum* JCM 9100 (4) *Halorubrum ejinorense* JCM 14265 (5) *Halorubrum lacusprofundi* ATCC 49239 (6) *Halorubrum lipolyticum* DSM 21995 (7) *Halorubrum litoreum* JCM 13561 (8) *Halorubrum saccharovorum* DSM 1137 (9) *Halorubrum sodomense* JCM 8880 (10) *Halorubrum tebenquichense* DSM 14210 (11) *Halorubrum terrestre* JCM 10247 (12) *Halorubrum tibetense* JCM 11889 (13) *Halorubrum trapanicum* JCM 10477 (14) *Halorubrum vacuolatum* JCM 9060 (15) *Halorubrum xinjiangense* JCM 12388 (16) *Halosarcina pallida* JCM 14848 (17) *Halosimplex carlsbadense* JCM 11222 (18) *Halostagnicola larsenii* JCM 13463.

have similar banding patterns according to the dendrogram analysis (e.g., *Natrinema ejinorense* and *Natrinema altunense*) but other species from the same genus are found elsewhere (e.g., *Natrinema pelliruberum* and *Natrinema versiforme*). This pattern is observed for all the genera for which several species were analyzed (e.g., *Halorubrum*, *Haloferax*). Thus, this DNA fingerprinting should not be used to classify isolates to a genus level. The observed amount of variation displayed among species within the same genus, led to the hypothesis that this technique might also detect genomic variation among strains within the same species. Therefore, we tested this fingerprinting technique on several populations of naturally co-occurring closely and distantly related strains.

MLSA ON ENVIRONMENTAL STRAINS

MLSA was performed in order to determine the genetic variation, and the evolutionary relationships of the isolates from Aran-Bidgol lake. Multiple sequence alignments were constructed from individual locus data from the new isolates and from genome data deposited in the NCBI database of type strains. Concatenated alignments were made from these and then a phylogenetic tree was constructed. The Aran-Bidgol isolates clustered into two main genera; *Halorubrum* and *Haloarcula* (**Figure 3**). Two polytomous groups, A and B, were observed within the genus *Halorubrum* and depicts evidence for distinct phylogroups with low sequence diversity as first seen for Spanish and Algerian isolates (Papke et al., 2007). Pairwise comparison of the number of nucleotides different within each of these phylogroups was carried out using MEGA 5 (Tamura et al., 2011). In both groups A and B, no two isolates had more than 10 nucleotide differences from one another across the concatenation of ~2500 bp (i.e., <1% sequence divergence; **Table 4**). This also holds true for group C (**Table 5**) within the *Haloarcula* cluster.

FINGERPRINTING THE ARAN-BIDGOL STRAINS

Genomic fingerprint analysis was run on each of the Aran-Bidgol lake environmental isolates. Banding patterns for each individual were generated and compared for similarity by dendrogram construction. The fingerprints and resulting dendrogram were then compared to the ML tree constructed from the MLSA data (**Figure 3**) for relating genetic and genomic variation within populations. It is noteworthy that despite limited numbers of bands produced for fingerprinting analysis, closely related strains from a single phylogroup displayed numerous variations in banding patterns, many of which were dissimilar to each other as determined by the dendrogram analysis. These widely different banding patterns reflect the variation in individual genomes. Comparison between sequence and banding pattern similarity demonstrates a lot of variation and no discernable patterns of relatedness even between strains that have zero differences across ~2500 nucleotides. Banding patterns of isolates within the genus *Halorubrum* seem as different as the banding patterns of isolates between the genera *Halorubrum* and *Haloarcula*. In some cases identical MLSA haplotypes have identical fingerprint patterns. We believe this can be attributed to the relatively low complexity of fingerprint bands produced, rather than two strains having identical genomes, and in such cases other methods of comparison like genome sequencing might reveal additional differences.

DISCUSSION

Our study employed DNA sequencing of multiple protein coding loci and random genomic amplification to test for variation in haloarchaeal isolates cultivated from the same location under the same conditions. The concatenated ML tree in **Figure 3**, and the number of pairwise nucleotide polymorphisms in **Tables 4**, **5**, show that many isolates are closely related to one another across

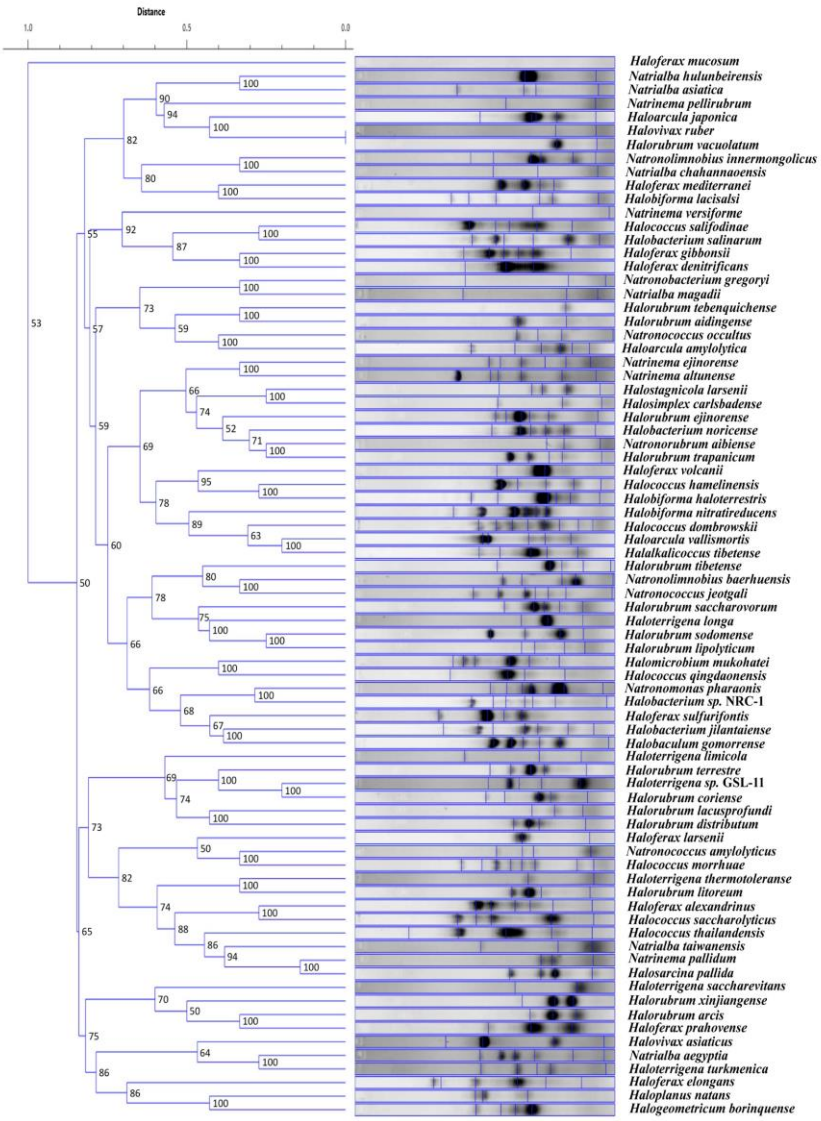


FIGURE 2 | UPGMA dendrogram comparing banding patterns between type strains. The numbers displayed at the nodes represent the cophenetic correlation coefficients.

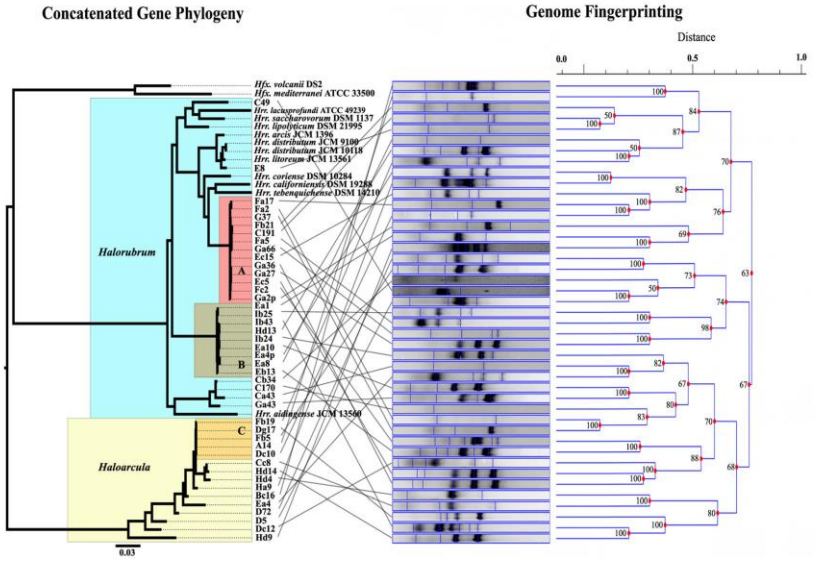


FIGURE 3 | Comparison of the maximum likelihood tree computed from the concatenation of five housekeeping genes, and the UPGMA dendrogram determined from the banding patterns of the genome fingerprinting. Lines between tree and dendrogram connect the same strain in the different analyses.

Table 4 | Pairwise comparison of number of nucleotide differences within polytomous Groups A and B defined on the maximum likelihood tree.

GROUP A	Ga27					0	7	8	5	7	10	9	6	Ea8	GROUP B
	Ec5	5					7	8	5	7	10	9	6	Ea4p	
	Ec15	8	5					5	2	8	7	6	5	Ea10	
	Ga66	7	8	7					5	9	8	7	4	Hd13	
	Fc2	5	4	5	6					6	9	8	3	Ib24	
	Fa2	1	1	1	0	1					5	4	7	Eb13	
	Fa5	7	8	7	4	6	0					1	8	Ib25	
	Fa17	2	2	2	0	2	0	1					7	Ea1	
	C191	8	9	8	5	7	0	1	1					Ib43	
	Fb21	8	9	8	5	7	0	1	1	0					
	Ga36	4	3	6	7	3	1	7	3	8	8				
	G37	8	3	4	7	5	0	5	1	6	6	6			
	Ga2p	6	5	4	5	3	0	5	1	6	6	4	4		

the five loci and are more or less indistinguishable from each other by these methods. However, the DNA fingerprinting analysis on these same isolates revealed additional variation not captured by MLSA, indicating genomic changes occur faster than the rate of substitution in redundant codon positions. Unfortunately, the deeper branches of the UPGMA hierarchical clustering dendrogram are unreliable for determining relationships and do not provide a good description of the measured Dice coefficients. Yet, shallower branches in the clustering diagram that are a good representation of the banding pattern differences show conflict with the MLSA phylogeny (**Figure 3**). Though the fingerprinting technique did not yield patterns of relatedness at the species level or genus level, it did demonstrate the high probability that the genomes of each isolates are unique. Whether that uniqueness is based on gene content or in genomic arrangements is undeterminable from this analysis.

However, given the known propensity for HGT in Halobacteria (Papke et al., 2004, 2007; Cuadros-Orellana

Table 5 | Pairwise comparison of number of nucleotide differences within polytomous Group C defined on the maximum likelihood tree.

GROUP C	Dc10							
	A14	0						
	Fb5	0	0					
	Dg17	0	0	0				
	Fb19	2	2	2	2			
	Hd14	30	30	30	30	32		
	Cc8	29	29	29	29	31	7	
	Hd4	38	38	38	38	39	26	19

Cells in blue represent members of Group C and cells in black represent the neighboring cluster on the ML tree.

et al., 2007; Lynch et al., 2012; Naor et al., 2012; Williams et al., 2012; Demaere et al., 2013; Podell et al., 2013), we surmise the fingerprint banding-pattern differences are largely due to gene transfer events. Discovery of recombinant hybrids (Naor et al., 2012) and the identification of enormous identical segments shared among the genomes of phylogenetically distant genera (Demaere et al., 2013) indicates the haloarchaea are subject to immense genomic variability from single gene transfer events. In another study, an influx of 303 transferred genes into *Haloferax mucosum* and *Haloferax mediterranei* were mostly of unknown function with some known transporters (Lynch et al., 2012), which is similar to the types of genes observed in the highly recombinogenic genomic islands of *Haloquadratum waslbyi* (Cuadros-Orellana et al., 2007). The *H. waslbyi* genome is 47.9% GC, but its genomic islands are GC rich by comparison, and enriched in transposable and repeat elements (Bolhuis et al., 2006) indicating a role for viruses in generating genomic diversity (Cuadros-Orellana et al., 2007). Similar to *H. walsbyi*, the genome of *Halobacterium* NRC-1 was interspersed with 91 insertion sequence elements of diverse GC compositions (Ng et al., 2000; Kennedy et al., 2001). Apart from HR, IS elements have been attributed to inactivating the bacterio-opsin gene in *Halobacterium halobium* (Dassarma et al., 1983) and causing genomic rearrangements at AT-rich regions in *Halobacterium* NRC-1 (Kennedy et al., 2001). Moreover, recent analysis indicates these Aran-Bidgol lake isolates display enormous variation in whole genome content with differences in group A ranging from 0.01 up to 0.51 Mb and from 0.07 up to 0.30 Mb in group B (Fullmer et al., 2014). Therefore, we hypothesize the drastic differences in fingerprints observed for the closest relatives (e.g., strains from groups A, B, and C) are more likely due to HGT, possibly mediated by insertion sequence elements (Dassarma et al., 1983; Ng et al., 2000; Kennedy et al., 2001), tRNAs (Naor et al., 2012), or other factors, rather than genome rearrangements.

We further suggest that the fingerprint banding patterns, especially for those within groups A, B, and C, were unlikely due to mutational events. Haloarchaea have low rates of spontaneous mutation, having been measured at 1.90×10^{-8} mutational events per cell division (Mackwan et al., 2007). Furthermore, haloarchaea are considered to have a high capacity for repairing DNA, as they have demonstrated the ability to survive radiation and desiccation damaged DNA (McCready, 1996; Kottemann et al., 2005), which is probably due to the prevalence of polyploidy through the process of gene conversion (Lange et al., 2011). Preliminary *in silico* analysis to determine the binding sites for each of the five primers in *Haloquadratum walsbyi* DSM 16790 and *Halorubrum lacusprofundi* revealed priming mostly in conserved loci, although a few phage related loci were also detected. Because many of the compared strains are very closely related, having only a few (or zero) nucleotide polymorphisms in the ~2500 sequenced base pairs, yet display enormous differences in fingerprint banding patterns, it would be unlikely that a few, or even one of the PCR binding sites in every strain within groups A, B, or C, would be mutated. Therefore, substitutions in PCR primer binding sites seem unlikely to have played a role in generating all the observed differences in banding patterns, especially those from closely related strains.

Analysis of five housekeeping genes demonstrates the isolates form genetically similar and distinct populations in a single environmental community and yet each genome is apparently different. This observation agrees well with expectations from the distributed genome hypothesis (Ehrlich et al., 2010). According to this, the non-core genes available in the pangenome pool are dispensed uniquely amongst the individual cells of a species. The differences in haloarchaeal genomic banding patterns suggests that in nature populations are made of highly varied individuals rather than clones of a single individual. The number of distinct genotypes observed, most likely due to gene flow, suggests that haloarchaeal cells are acquiring genomic variation within populations at a rate faster than redundant codon position substitutions, and possibly at every replication event. Distribution of the non-core genes within a highly recombining population (defined by MLSA phylogeny) theoretically enables the individual to quickly adapt to new environmental selection conditions, especially virus predation (Cuadros-Orellana et al., 2007) but may also result from random processes like neutral drift (Gogarten and Townsend, 2005).

AUTHOR CONTRIBUTIONS

R. Thane Papke, J. Peter Gogarten, and Antonio Ventosa conceived the researched. Nikhil Ram Mohan, Matthew S. Fullmer, Andrea M. Makkay, and Ryan Wheeler gathered data, and performed the analyses. Nikhil Ram Mohan, Matthew S. Fullmer, Andrea M. Makkay, Ryan Wheeler, Antonio Ventosa, J. Peter Gogarten, and R. Thane Papke wrote the manuscript.

ACKNOWLEDGMENTS

We would like to thank Mohammad A. Amozegar from the University of Tehran for cultivating the Aran-Bidgol salt lake isolates. This research was supported by the National Science Foundation (award numbers, DEB0919290 and DEB0830024) and NASA Astrobiology: Exobiology and Evolutionary Biology Program Element (Grant Number NNX12AD70G).

REFERENCES

Akaike, H. (1974). A new look at the statistical model. *IEEE Trans. Autom. Control* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Allers, T., Ngo, H. P., Mevarech, M., and Lloyd, R. G. (2004). Development of additional selectable markers for the halophilic archaeon *Haloferax volcanii* based on the *leuB* and *trpA* genes. *Appl. Environ. Microbiol.* 70, 943–953. doi: 10.1128/AEM.70.2.943-953.2004

Andam, C. P., Harlow, T. J., Papke, R. T., and Gogarten, J. P. (2012). Ancient origin of the divergent forms of leucyl-tRNA synthetases in the *Halobacteriales*. *BMC Evol. Biol.* 12:85. doi: 10.1186/1471-2148-12-85

Anton, J., Llobet-Brossa, E., Rodriguez-Valera, F., and Amann, R. (1999). Fluorescence *in situ* hybridization analysis of the prokaryotic community inhabiting crystallizer ponds. *Environ. Microbiol.* 1, 517–523. doi: 10.1046/j.1462-2920.1999.00065.x

Barrangou, R., Yoon, S. S., Breidt Jr. F. Jr., Fleming, H. P., and Klaenhammer, T. R. (2002). Characterization of six *Leuconostoc fallax* bacteriophages isolated from an industrial sauerkraut fermentation. *Appl. Environ. Microbiol.* 68, 5452–5458. doi: 10.1128/AEM.68.11.5452-5458.2002

Bolhuis, H., Palm, P., Wende, A., Falb, M., Rampp, M., Rodriguez-Valera, F., et al. (2006). The genome of the square archaeon *Haloquadratum walsbyi*: life at the limits of water activity. *BMC Genomics* 7:169. doi: 10.1186/1471-2164-7-169

Boucher, Y., Douady, C. J., Sharma, A. K., Kamekura, M., and Doolittle, W. F. (2004). Intragenomic heterogeneity and intergenomic recombination among haloarchaeal rRNA genes. *J. Bacteriol.* 186, 3980–3990. doi: 10.1128/JB.186.12.3980-3990.2004

Cuadros-Orellana, S., Martin-Cuadrado, A. B., Legault, B., D'Auria, G., Zhaxybayeva, O., Papke, R. T., et al. (2007). Genomic plasticity in prokaryotes: the case of the square haloarchaeon. *ISME J.* 1, 235–245. doi: 10.1038/ismej.2007.35

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* 9, 772. doi: 10.1038/nmeth.2109

Dassarma, S., Rajbhandary, U. L., and Khorana, H. G. (1983). High-frequency spontaneous mutation in the bacterio-opsin gene in *Halobacterium halobium* is mediated by transposable elements. *Proc. Natl. Acad. Sci. U.S.A.* 80, 2201–2205. doi: 10.1073/pnas.80.8.2201

De la Haba, R. R., Marquez, M. C., Papke, R. T., and Ventosa, A. (2012). Multilocus sequence analysis of the family *Halomonadaceae*. *Int. J. Syst. Evol. Microbiol.* 62, 520–538. doi: 10.1099/ijs.0.032938-0

Demaere, M. Z., Williams, T. J., Allen, M. A., Brown, M. V., Gibson, J. A., Rich, J., et al. (2013). High level of intergenera gene exchange shapes the evolution of haloarchaea in an isolated antarctic lake. *Proc. Natl. Acad. Sci. U.S.A.* 110, 16939–16944. doi: 10.1073/pnas.1307090110

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* 26, 6. doi: 10.2307/1932409

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Ehrlich, G. D., Ahmed, A., Earl, J., Hiller, N. L., Costerton, J. W., Stoodley, P., et al. (2010). The distributed genome hypothesis as a rubric for understanding evolution *in situ* during chronic bacterial biofilm infectious processes. *FEMS Immunol. Med. Microbiol.* 59, 269–279. doi: 10.1111/j.1574-695X.2010.00704.x

Feil, E. J., Enright, M. C., and Spratt, B. G. (2000). Estimating the relative contributions of mutation and recombination to clonal diversification: a comparison between *Neisseria meningitidis* and *Streptococcus pneumoniae*. *Res. Microbiol.* 151, 465–469. doi: 10.1016/S0923-2508(00)00168-6

Fullmer, M. S., Soucy, S. M., Swithers, K. S., Makkay, A. M., Wheeler, R., Ventosa, A., et al. (2014). Population and genomic analysis of the genus *Halorubrum*. *Front. Microbiol.* 5:140. doi: 10.3389/fmicb.2014.00140

Ghai, R., Pasic, L., Fernandez, A. B., Martin-Cuadrado, A. B., Mizuno, C. M., McMahon, K. D., et al. (2011). New abundant microbial groups in aquatic hypersaline environments. *Sci. Rep.* 1:135. doi: 10.1038/srep00135

Gogarten, J. P., and Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.* 3, 679–687. doi: 10.1038/nrmicro1204

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Jones, J. G., Young, D. C., and Dassarma, S. (1991). Structure and organization of the gas vesicle gene cluster on the *Halobacterium halobium* plasmid pNRC100. *Gene* 102, 117–122. doi: 10.1016/0378-1119(91)90549-Q

Kennedy, S. P., Ng, W. V., Salzberg, S. L., Hood, L., and Dassarma, S. (2001). Understanding the adaptation of *Halobacterium* species NRC-1 to its extreme environment through computational analysis of its genome sequence. *Genome Res.* 11, 1641–1650. doi: 10.1101/gr.190201

Kottemann, M., Kish, A., Iloanusi, C., Bjork, S., and Diruggiero, J. (2005). Physiological responses of the halophilic archaeon *Halobacterium* sp. strain NRC1 to desiccation and gamma irradiation. *Extremophiles* 9, 219–227. doi: 10.1007/s00792-005-0437-4

Lange, C., Zerulla, K., Breuert, S., and Soppa, J. (2011). Gene conversion results in the equalization of genome copies in the polyploid haloarchaeon *Haloferax volcanii*. *Mol. Microbiol.* 80, 666–677. doi: 10.1111/j.1365-2958.2011.07600.x

Lapierre, P., and Gogarten, J. P. (2009). Estimating the size of the bacterial pan-genome. *Trends Genet.* 25, 107–110. doi: 10.1016/j.tig.2008.12.004

Legault, B. A., Lopez-Lopez, A., Alba-Casado, J. C., Doolittle, W. F., Bolhuis, H., Rodriguez-Valera, F., et al. (2006). Environmental genomics of "*Haloquadratum walsbyi*" in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. *BMC Genomics* 7:171. doi: 10.1186/1471-2164-7-171

Litchfield, C. D. (1998). Survival strategies for microorganisms in hypersaline environments and their relevance to life on early Mars. *Meteorit. Planet. Sci.* 33, 813–819. doi: 10.1111/j.1945-5100.1998.tb01688.x

Lozier, R. H., Bogomolni, R. A., and Stoeckenius, W. (1975). Bacteriorhodopsin: a light-driven proton pump in *Halobacterium halobium*. *Biophys. J.* 15, 955–962. doi: 10.1016/S0006-3495(75)85875-9

Lynch, E. A., Langille, M. G., Darling, A., Wilbanks, E. G., Haltiner, C., Shao, K. S., et al. (2012). Sequencing of seven haloarchaeal genomes reveals patterns of genomic flux. *PLoS ONE* 7:e41389. doi: 10.1371/journal.pone.0041389

Mackwan, R. R., Carver, G. T., Drake, J. W., and Grogan, D. W. (2007). An unusual pattern of spontaneous mutations recovered in the halophilic archaeon *Haloferax volcanii*. *Genetics* 176, 697–702. doi: 10.1534/genetics.106.069666

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U.S.A.* 95, 3140–3145. doi: 10.1073/pnas.95.6.3140

Martinez-Murcia, A. J., and Rodriguez-Valera, F. (1994). The use of arbitrarily primed PCR (AP-PCR) to develop taxa specific DNA probes of known sequence. *FEMS Microbiol. Lett.* 124, 265–270. doi: 10.1016/0378-1097(94)00440-4

McCready, S. (1996). The repair of ultraviolet light-induced DNA damage in the halophilic archaebacteria, *Halobacterium cutirubrum*, *Halobacterium halobium* and *Haloferax volcanii*. *Mutat. Res.* 364, 25–32. doi: 10.1016/0921-8777(96)00018-3

Naor, A., Lapierre, P., Mevarech, M., Papke, R. T., and Gophna, U. (2012). Low species barriers in halophilic archaea and the formation of recombinant hybrids. *Curr. Biol.* 22, 1444–1448. doi: 10.1016/j.cub.2012.05.056

Nelson-Sathi, S., Dagan, T., Landan, G., Janssen, A., Steel, M., McInerney, J. O., et al. (2012). Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc. Natl. Acad. Sci. U.S.A.* 109, 20537–20542. doi: 10.1073/pnas.1209119109

Ng, W. V., Kennedy, S. P., Mahairas, G. G., Berquist, B., Pan, M., Shukla, H. D., et al. (2000). Genome sequence of *Halobacterium* species NRC-1. *Proc. Natl. Acad. Sci. U.S.A.* 97, 12176–12181. doi: 10.1073/pnas.190337797

Oren, A. (2002). Molecular ecology of extremely halophilic archaea and bacteria. *FEMS Microbiol. Ecol.* 39, 1–7. doi: 10.1111/j.1574-6941.2002.tb00900.x

Papke, R. T., Koenig, J. E., Rodriguez-Valera, F., and Doolittle, W. F. (2004). Frequent recombination in a saltern population of *Halorubrum*. *Science* 306, 1928–1929. doi: 10.1126/science.1103289

Papke, R. T., White, E., Reddy, P., Weigel, G., Kamekura, M., Minegishi, H., et al. (2011). A multilocus sequence analysis approach to the phylogeny and taxonomy of the *Halobacteriales*. *Int. J. Syst. Evol. Microbiol.* 61, 2984–2995. doi: 10.1099/ijs.0.029298-0

Papke, R. T., Zhaxybayeva, O., Feil, E. J., Sommerfeld, K., Muise, D., and Doolittle, W. F. (2007). Searching for species in haloarchaea. *Proc. Natl. Acad. Sci. U.S.A.* 104, 14092–14097. doi: 10.1073/pnas.0706358104

Podell, S., Ugalde, J. A., Narasingarao, P., Banfield, J. F., Heidelberg, K. B., and Allen, E. E. (2013). Assembly-driven community genomics of a hypersaline microbial ecosystem. *PLoS ONE* 8:e61692. doi: 10.1371/journal.pone.0061692

Sharma, A. K., Walsh, D. A., Bapteste, E., Rodriguez-Valera, F., Doolittle, W. F., and Papke, R. T. (2007). Evolution of rhodopsin ion pumps in haloarchaea. *BMC Evol. Biol.* 7:79. doi: 10.1186/1471-2148-7-79

Shivu, M. M., Rajeeva, B. C., Girisha, S. K., Karunasagar, I., and Krohne, G. (2007). Molecular characterization of *Vibrio harveyi* bacteriophages isolated from aquaculture environments along the coast of India. *Environ. Microbiol.* 9, 322–331. doi: 10.1111/j.1462-2920.2006.01140.x

Sokal, R. R., and Rohlf, F. J. (1962). The Comparison of dendrograms by objective methods. *Taxon* 11, 8. doi: 10.2307/1217208

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739. doi: 10.1093/molbev/msr121

Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* 11, 472–477. doi: 10.1016/j.mib.2008.09.006

Thompson, J. R., Pacocha, S., Pharino, C., Klepac-Ceraj, V., Hunt, D. E., Benoit, J., et al. (2005). Genotypic diversity within a natural coastal bacterioplankton population. *Science* 307, 1311–1313. doi: 10.1126/science.1106028

Williams, D., Gogarten, J. P., and Papke, R. T. (2012). Quantifying homologous replacement of loci between haloarchaeal species. *Genome Biol. Evol.* 4, 1223–1244. doi: 10.1093/gbe/evs098

Winget, D. M., and Wommack, K. E. (2008). Randomly amplified polymorphic DNA PCR as a tool for assessment of marine viral richness. *Appl. Environ. Microbiol.* 74, 2612–2618. doi: 10.1128/AEM.02829-07

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 07 February 2014; accepted: 19 March 2014; published online: 09 April 2014.
Citation: Ram Mohan N, Fullmer MS, Makkay AM, Wheeler R, Ventosa A, Naor A, Gogarten JP and Papke RT (2014) Evidence from phylogenetic and genome fingerprinting analyses suggests rapidly changing variation in Halorubrum and Haloarcula populations. Front. Microbiol. 5:143. doi: 10.3389/fmicb.2014.00143
This article was submitted to Extreme Microbiology, a section of the journal Frontiers in Microbiology.

Appendix B – Soucy et al., 2014

Inteins as indicators of gene flow in the halobacteria

frontiers in
MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 26 June 2014
doi: 10.3389/fmicb.2014.00299

Inteins as indicators of gene flow in the halobacteria

*Shannon M. Soucy, Matthew S. Fullmer, R. Thane Papke and Johann Peter Gogarten**

Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA

Edited by:
Jesse Dillon, California State University, Long Beach, USA

Reviewed by:
Julie L. Meyer, University of Florida, USA
Kenneth Mills, College of the Holy Cross, USA

**Correspondence:*
Johann Peter Gogarten, Microbiology Program, Department of Molecular and Cell Biology, University of Connecticut, 91 N. Eagleville Rd., Storrs, CT 06269-3125, USA
e-mail: gogarten@uconn.edu; jpgogarten@gmail.com

This research uses inteins, a type of mobile genetic element, to infer patterns of gene transfer within the Halobacteria. We surveyed 118 genomes representing 26 genera of Halobacteria for intein sequences. We then used the presence-absence profile, sequence similarity and phylogenies from the inteins recovered to explore how intein distribution can provide insight on the dynamics of gene flow between closely related and divergent organisms. We identified 24 proteins in the Halobacteria that have been invaded by inteins at some point in their evolutionary history, including two proteins not previously reported to contain an intein. Furthermore, the size of an intein is used as a heuristic for the phase of the intein's life cycle. Larger size inteins are assumed to be the canonical two domain inteins, consisting of self-splicing and homing endonuclease domains (HEN); smaller sizes are assumed to have lost the HEN domain. For many halobacterial groups the consensus phylogenetic signal derived from intein sequences is compatible with vertical inheritance or with a strong gene transfer bias creating these clusters. Regardless, the coexistence of intein-free and intein-containing alleles reveal ongoing transfer and loss of inteins within these groups. Inteins were frequently shared with other Euryarchaeota and among the Bacteria, with members of the Cyanobacteria (*Cyanothece, Anabaena*), Bacteriodetes (*Salinibacter*), Betaproteobacteria (*Delftia, Acidovorax*), Firmicutes (*Halanaerobium*), Actinobacteria (*Longispora*), and Deinococcus-Thermus-group.

Keywords: gene symbiosis, genome as an ecosystem, inteins, mobile genetic elements, gene flow, horizontal gene transfer, halobacteria

INTRODUCTION

Inteins are self-splicing genetic parasites located in highly conserved sites of slowly evolving genes. They are found in all three domains of life and in viruses (Perler et al., 1997; Pietrokovski, 2001; Gogarten et al., 2002; Swithers et al., 2009). Similar to group I introns, inteins are often associated with a homing endonuclease (HEN). An important difference between inteins and introns is the timing of the splicing activity, which occurs immediately after transcription in introns and after translation in inteins (Hirata et al., 1990; Kane et al., 1990). The association with a HEN domain enables a cyclic invasion pattern, called the homing cycle (Goddard and Burt, 1999; Gogarten and Hilario, 2006). The homing cycle consists of three phases: intein invasion, intein fixation, and eventually loss of the intein enabling invasion to occur again. During invasion and fixation the intein splicing domains are associated with a HEN domain forming a canonical intein (hereafter referred to as a large intein); however, during the loss phase the function of the HEN is often disrupted and begins to degrade, generating a mini-intein. Simulations have shown that intein-containing and intein-free alleles can coexist in well mixed populations under some sets of parameters (Pahara et al., 2009; Barzel et al., 2011). Also, inteins with functioning HEN domains were inferred to have persisted in some eukaryotic lineages for several 100 million years (Butler et al., 2006; Gogarten and Hilario, 2006).

Inteins do not have an apparatus to penetrate the cell envelope. Therefore, they must rely on mechanisms in place within the population for insertion into the cell such as: conjugation, mating, generalized DNA uptake, and viruses or gene transfer agents (Lang et al., 2012). The faster-than-Mendelian inheritance of the large inteins (Gimble and Thorner, 1992), along with a nearly neutral fitness burden, enables these mobile elements to persist in organisms over evolutionary time as long as there are new populations to invade (Goddard and Burt, 1999; Gogarten and Hilario, 2006). Furthermore, the size of the intein (mini or large) provides information about the genomic mobility of the element as mini inteins are rarely integrated into the recipient's genome; whereas large inteins are more frequently integrated due to the activity of the HEN. The conservation of the recognition site provides an invasion target even in distantly related strains and species. Also, inteins have a higher substitution rate relative to their extein hosts (Swithers et al., 2013). This substitution rate gives rise to many evolutionarily informative sites when comparing a large collection of homologous inteins. In this work, we take advantage of these traits and survey the distribution of inteins in the Halobacteria, a highly recombinant class of halophilic Archaea (Williams et al., 2012) known to contain several intein alleles (Perler, 2002). We make use of 118 halobacterial genomes (Supplementary Table 1) and the previously reported and newly discovered intein alleles to survey networks of gene transfer within and outside the

Halobacteria based on the presence-absence profile of the inteins, their sequence similarity, and the phylogenies reconstructed from intein sequences.

MATERIALS AND METHODS

HALOBACTERIAL INTEIN SEQUENCE RETRIEVAL AND ALIGNMENT

Position specific scoring matrices (PSSMs) were created using the collection of all inteins from InBase, the Intein database and registry (Perler, 2002). A custom database was created with all inteins, and each intein was used as a seed to create a PSSM using the custom database. These PSSMs were then used as a seed for PSI-BLAST (Altschul, 1997) searches against each of the halobacterial genomes available from NCBI as of June 2013 as well as a private collection sequenced by our collaborators. To remove false positives, a size exclusion step was then performed on each protein sequence as an intein domain adds 100–700 aa to invaded protein sequences. Inteins were then aligned using Muscle (Edgar, 2004) with default parameters in the SeaView version 4.0 software package (Gouy et al., 2010). Insertions, which passed the size exclusion step, but did not contain splicing domains, were removed and the previous steps were repeated using the resulting dataset on a collection of private genomes from the Papke lab. Prottest 3.2 (Guindon et al., 2010; Darriba et al., 2011) was used to determine an appropriate substitution model for the intein sequences, the WAG model was favored and used for all subsequent trees for consistency. Once the collection of halobacterial inteins was complete, sequences were re-aligned using SATé (Liu et al., 2012) to generate a final alignment using MAFFT (Katoh and Standley, 2013) to align, Muscle (Edgar, 2004) to merge, RAXML(Stamatakis, 2014) for tree estimation, and a WAG model for each allele.

To determine the relationship among all halobacterial inteins, the inteins were aligned using Muscle (Edgar, 2004). Subsequently a tree was built using PhyML v3.0 (Guindon et al., 2010) using a WAG substitution model with a Gamma shape parameter and the proportion of invariant sites estimated from the data.

INTEIN RETRIEVAL OUTSIDE THE HALOBACTERIA

Each halobacterial intein was used as a BLAST (Altschul et al., 1990) query against the non-redundant database on NCBI. Any match with an e-value better than 0.000001 was aligned to the dataset to which its query belonged. Sequences were then filtered based on the protein annotation and goodness of fit to the existing alignment. As an additional filtering step each match was used as a query against the non-redundant database and the majority BLAST hit annotations were used to verify the protein identity, as annotations are not always reliable. Remaining sequences were aligned using Clustal Omega 1.1.0 (Sievers et al., 2011) with the profile alignment option in SeaView 4.0 (Gouy et al., 2010). Maximum-likelihood trees were built using PhyML (Guindon et al., 2010) with the WAG model, and rates estimated from the data.

To assess the relative contribution of different genera represented in each intein allele sequence data set, a stacked column graph was created. Sequence density was calculated for each intein allele by dividing the number of intein sequences in each genus by the number of total intein sequences in that allele.

SYMBIOTIC STATE ASSIGNMENT

Intein sequence length was used to determine symbiotic state. For each intein allele the length of the intein sequence was determined. A cutoff length for mini-intein assignment was based on the presence of a gap in intein lengths greater than 100 amino acids within an allele. The third intein state "no-intein" was assigned where the intein was clearly absent from the orthologous protein containing an intein in any of the halobacterial genomes examined. Additionally, once an intein was noted as a mini-intein the alignment was analyzed to ensure the gaps in these sequences correspond to the location of the HEN domain.

RIBOSOMAL PROTEIN REFERENCE TREE

Alignments of 55 ribosomal protein for 21 Halobacteria (Williams et al., 2012) were used to find orthologous proteins in the genomes used in this work. In-house python scripts (data file 1) were used to concatenate the alignments, and PhyML v3.0 (Guindon et al., 2010) was used to build a tree. The tree used the WAG substitution model with the Gamma shape parameter and the proportion of invariant sites and base frequencies estimated from the data.

BAYESIAN CLUSTERING WITH INTEIN SEQUENCES

A concatenation of an intein presence-absence matrix and alignments for each intein allele were generated using in-house python scripts (data file 1). MrBayes version 3.2.1 (Ronquist et al., 2012) was then used to perform a clustering analysis using a partition allowing for character states in the presence-absence matrix and sequence information for each intein allele. The prior for the character portion of the data matrix used a symmetrical Dirichlet distribution with an exponential (1.0), and variable rates so each column was considered independent of the others. The likelihood for the character portion of the alignment used variable coding and 5 beta categories. The prior for the protein sequences in the alignment used a fixed WAG substitution model, with state frequencies estimated from the data, and the likelihood settings used a Gamma shape parameter and the proportion of invariant sites estimated from the data.

RESULTS

HALOBACTERIAL INTEINS

The intein content of a collection of halobacterial genomes was analyzed using an intein-allele-specific PSSM. This survey revealed 13 genes in the Halobacteria invaded by inteins at 24 distinct positions (intein alleles) (Table 1). Seven of these intein alleles were not previously reported in the Halobacteria, and two of the seven have not previously been reported to harbor inteins: a DNA ligase gene involved in double strand break repair, and a deaminase gene involved in nucleotide metabolism (Table 1). To determine if vertical inheritance was accountable for the distribution of intein alleles, the presence–absence matrix of intein alleles was mapped onto a reference phylogeny (Figure 1). Clearly, intein presence-absence is not concordant with the ribosomal protein phylogeny, implicating abundant horizontal genetic transfer (HGT) in creating the observed distribution. The presence of multiple intein alleles in the majority of genomes (70%) might be interpreted

Table 1 | Exteins in the halobacteria.

Intein allele	Extein annotation
cdc21-a	Cell division control protein 21
cdc21-b	
cdc21-c	
polB-d	DNA polymerase B1
polB-a	
polB-b	
polB-c	
pol-IIa	DNA polymerase II large subunit
pol-IIb#	
dtd##	Deoxycytadine triphosphate deaminase
gyrB	DNA gyrase subunit B
helicase-b#	ATP-dependent helicase
ligase##	ATP-dependent DNA ligase I
rfc-a	Replication factor C small subunit
rfc-d#	
rir1-l#	Ribonucleoside-diphosphate reductase
rir1-k	
rir1-b	
rir1-g	
rir1-m#	
rpolA	DNA-directed RNA polymerase subunit A
udp	UDP-glucose 6-dehydrogenase
topA	DNA topoisomerase I
top6B	DNA topoisomerase VI subunit B

#*Denotes intein alleles discovered in this work.*

##*Denotes extein sequences not previously reported to be invaded by an intein.*

to suggest that inteins could spread locally within a single genome.

INTEIN PROPAGATION WITHIN THE HALOBACTERIA

To address the possibility of inteins moving locally within a genome, the phylogenetic relationships among all halobacterial intein sequences were analyzed (Figure 2). All of the intein alleles form highly supported clusters with others of the same type, with the exception of two sequences: the polB-c inteins of Haloferax larsenii and Haloferax elongans group inside the polB-b intein allele cluster; however, this node is poorly supported (59/100 bootstraps) indicating this relationship could be an artifact produced by poor resolution of the relationships that connect various intein alleles. Furthermore, there is poor support linking all of the intein allele clusters together (less than 70% bootstrap support), indicating sequence conversion (an intein invading an ectopic or atypical locus) between intein alleles, even within the same host protein, is uncommon. Among the inteins analyzed here, at most one invasion of an ectopic site is supported by the data, confirming that this type of event is rare (Perler et al., 1997; Gogarten et al., 2002). These data indicate that HGT is the only plausible explanation for the large number of different intein alleles in this class of organisms. Incongruence between the presence of inteins and ribosomal phylogeny also support this conclusion.

BAYESIAN PHYLOGENETIC ANALYSIS OF INTEINS

In an attempt to resolve the local events (transfers and vertical inheritance within the Halobacteria) that gave rise to the observed intein distribution in the Halobacteria, a Bayesian analysis based on the intein sequences for each allele and on the presence-absence pattern was performed (Figure 3). In this analysis two organisms may group together because they both inherited inteins from a common ancestor, or because an intein was recently transferred between them. The paucity of well-supported nodes (nodes with 0.95 or greater posterior probability were considered well-supported) in part reflects the extent to which our sample is biased toward very similar sequences (31% of halobacterial genomes in this study are from Halorubrum). Most of the well-supported clusters in the Bayesian tree also occur in the reference tree, suggesting these inteins may be the result of shared vertical inheritance. However, many of these clusters do not have identical intein profiles (clusters 1, 6, 8, and 10), thus HGT between close relatives is a better explanation than vertical inheritance for these clusters. Only three of the clusters, 2, 9, and 12, have branching orders that are different from those observed in the reference tree indicating HGT. Cluster 2 is made up of Natrinema spp. pellirubrum and versiforme which share only the pol-IIa intein. In the reference tree Nnm. versiforme groups with the rest of the Natrinema, and Nnm. pellirubrum groups with Haloterrigena thermotolerans. Natrinema sp. J7-2 is the only other member of the Natrinema that has an intein in the pol-IIa position, but the intein in this species is 14 aa shorter than the intein shared by Nnm. pellirubrum and Nnm. versiforme. Htg. thermotolerans shares no inteins with Nnm. pellirubrum. Cluster 9 is made up of Halorubrum spp. C49 and E3, which share only the cdc21-b intein. In the reference tree Hrr. E3 groups with Halorubrum litoreum and the two share the pol-IIa intein allele, but no others. Hrr. C49 groups with Halorubrum saccharovorum and they do not share any inteins. Cluster 12 is made up of Haloferax spp. denitrificans, lucentense, alexandrinus, and Haloferax sp. BAB2207, which all have an intein in the cdc21-a position. In the reference tree Hfx. lucentense, Hfx. sp. BAB2207, and Hfx. alexandrinus all group together, but Hfx. denitrificans groups with Haloferax sulfurifontis, and they do not share any inteins. The lack of shared inteins between clusters in the reference tree and differences among the inteins shared in these clusters cause these divergences in this tree as compared to the reference tree. This may indicate that the taxa in the Bayesian clusters are exchange partners, or that they share unsampled intermediate exchange partners. Additionally, the majority of clusters share 2 or fewer intein alleles between all members of the cluster (eight out of 12 clusters). The two clusters that share the most intein alleles between all members are Cluster 3, made up of Haloqudratum walsbyi strains DSM 16790 and C23 with 13 shared intein alleles, and cluster 7 made up of Halorubrum spp. strains SP3 and SP9 sharing 4 intein alleles. Both of these clusters have branching patterns identical to those on the reference tree, indicating that phylogenetic proximity plays a significant role in intein distribution.

Members of the Halorubrum genus, not surprisingly, were highly represented in the clusters (four of 12 total). All four of the clusters show a geographic bias. Clusters 6, 8, and 9 were all isolated from the Aran-Bidgol lake in Iran, and cluster 7 was isolated from the Sedom Ponds in Israel (Atanasova et al., 2012). Branch lengths in all of these clusters are very small, suggesting these populations are well mixed with respect to intein sequences. Geography does not seem to play a strong role in linking other

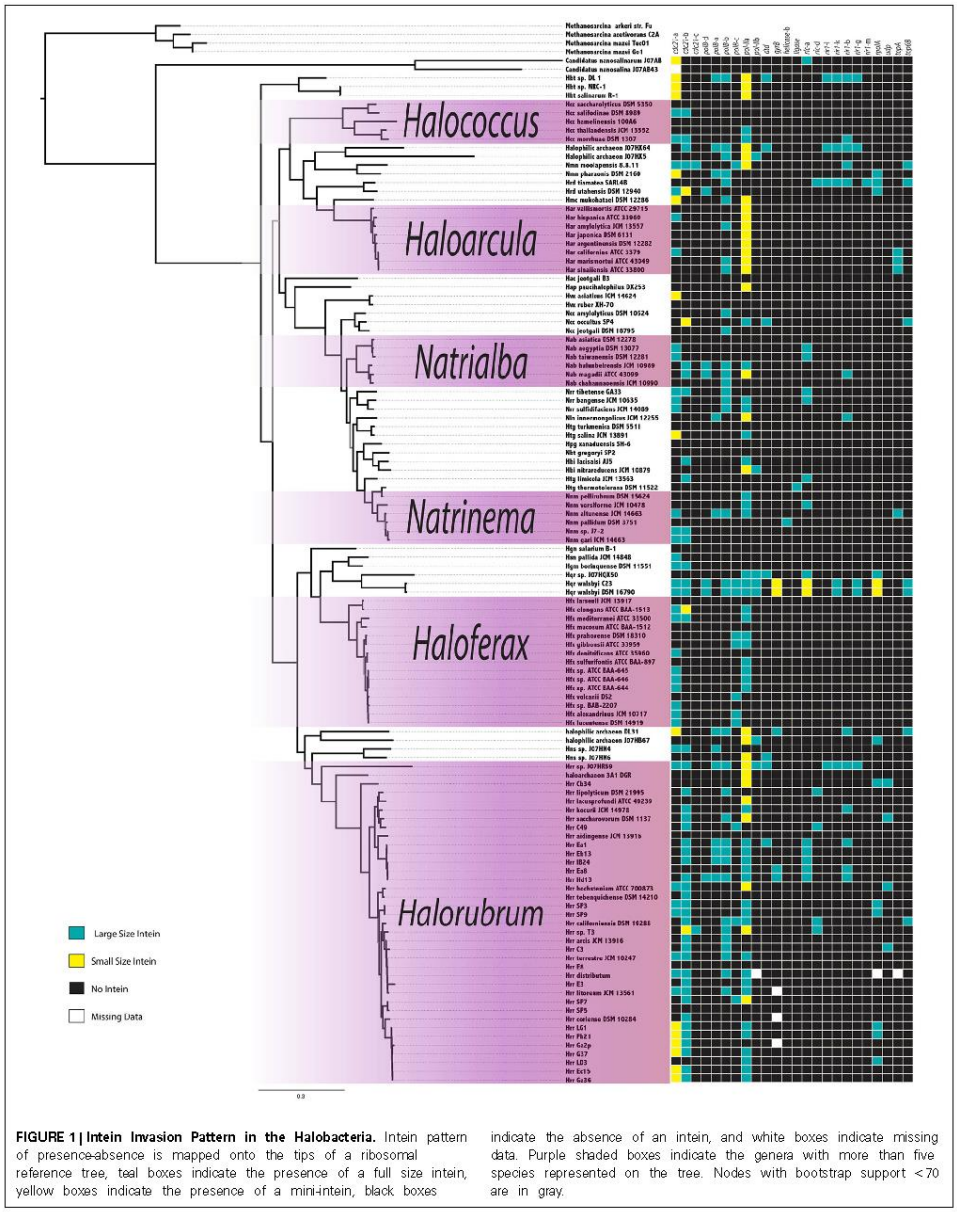


FIGURE 1 | Intein Invasion Pattern in the Halobacteria. Intein pattern of presence-absence is mapped onto the tips of a ribosomal reference tree, teal boxes indicate the presence of a full size intein, yellow boxes indicate the presence of a mini-intein, black boxes indicate the absence of an intein, and white boxes indicate missing data. Purple shaded boxes indicate the genera with more than five species represented on the tree. Nodes with bootstrap support < 70 are in gray.

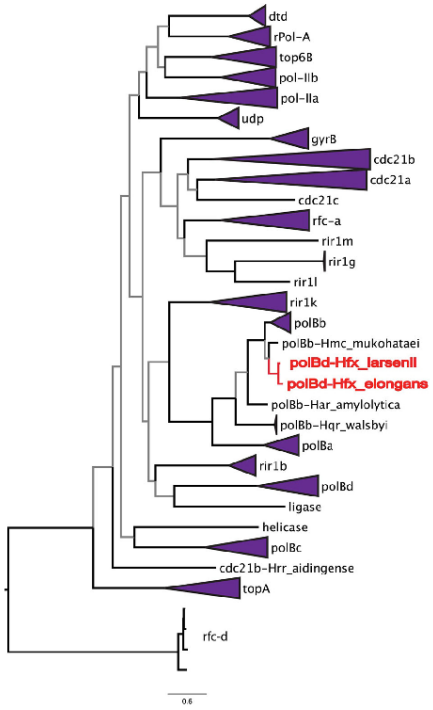


FIGURE 2 | Relationships among intein alleles in the Halobacteria. This tree depicts the phylogenetic relationships among intein alleles in the Halobacteria. Inteins that clustered in concordance with the allele were collapsed to a single node, and labeled with the name of the intein allele. Two *polB*-d sequences did not group with the *polB*-d allele, and instead are located amongst the *polB*-c alleles these are indicated in red. Nodes with bootstrap support < 70 are colored gray.

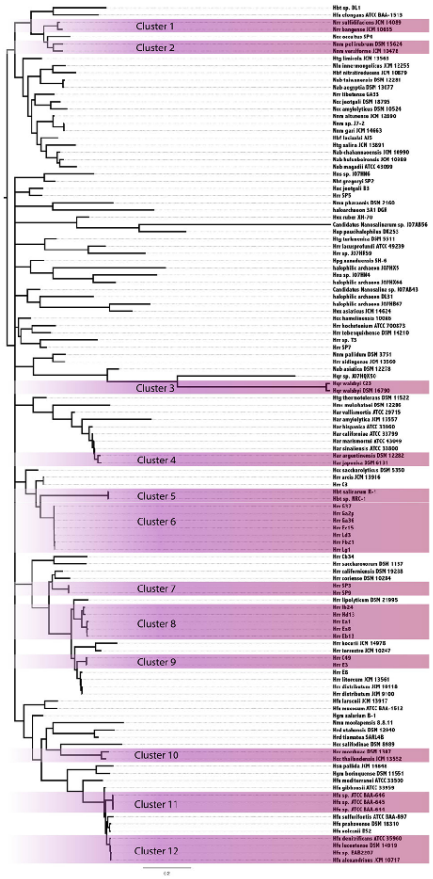


FIGURE 3 | Clustering of Halobacteria based on intein sequences and distribution. Halobacteria were clustered based on intein sequences and the distribution in each genome. Clusters with posterior probability > 95% are shaded purple.

well-supported clusters based on intein sequences. Furthermore, evidence of clustering based on geography in the **Halorubrum** is less interesting than the clear separation between groups isolated from the same location (cluster 6, 8, and 9). This separation of species of **Halorubrum** from the same location is echoed in the reference tree, and taken together with the short branch lengths in these clusters indicate that population structure plays a strong role in gene sharing at least for this location (see Fullmer et al., 2014 for in depth discussion). Increased geographical sampling could reveal similar trends in other locations.

INTEIN HOMING IN THE HALOBACTERIA

The existence of a singleton in an intein allele in the genomes analyzed could represent intein invasion from outside the Halobacteria; but could also be due to incomplete sampling. To investigate the phylogenetic distance of invasion events responsible for the observed distribution of inteins, the halobacterial inteins were used as queries to search for homologous sequences in the non-redundant database (Altschul et al., 1990). Intein sequences that matched the alleles in the Halobacteria were found in other Euryarchaeota (but not Crenarchaeota), and Bacteria (**Table 2**). To ascertain whether homing occurred between the Halobacteria and organisms outside the Halobacteria, a maximum likelihood tree was built for each intein allele. The

Table 2 | Taxonomic distribution in each intein allele.

Intein allele	Tree topology	Halobacteria	Bacteria	Other Euryarchaeota
cdc21-a	Monophyletic	55	4	16
cdc21-c	Monophyletic	1	0	0
*dtd***	Monophyletic	6	0	0
gyrB	Monophyletic	6	19	1
*helicase-b**	Monophyletic	1	2	1
*ligase***	Monophyletic	1	0	0
*pol-IIb**	Monophyletic	9	0	1
polB-d	Monophyletic	6	0	1
rfc-a	Monophyletic	16	0	13
*rfc-d**	Monophyletic	5	0	0
rir1-b	Monophyletic	15	55	5
rir1-g	Monophyletic	4	15	0
rir1-k	Monophyletic	5	1	0
*rir1-l**	Monophyletic	3	3	0
rpolA	Monophyletic	10	0	0
top6B	Monophyletic	8	0	0
topA	Monophyletic	4	0	1
udp	Monophyletic	7	2	6
*rir1-m**	Monophyletic	1	4	0
polB-c	Monophyletic	20	1	1
polB-a	Polyphyletic-bacteria	16	2	1
polB-b	Polyphyletic-bacteria	38	3	0
pol-IIa	Polyphyletic-Euryarchaeota	75	0	16
cdc21-b	Polyphyletic-Euryarchaeota	51	1	3

*Denotes intein alleles discovered in this work.
**Denotes exteins discovered in this work.

tree topologies were evaluated with respect to the halobacterial inteins. If the halobacterial inteins in the tree were monophyletic it was assumed that except for the initial invasion gene flow for that intein allele occurred within the Halobacteria exclusively. If the halobacterial inteins were polyphyletic, invasion events that generated the observed distribution likely involved organisms outside the Halobacteria either as donors or as recipients. The majority of intein trees, 83%, were monophyletic, reinforcing the idea that recombination is more successful between closely related organisms (Gogarten et al., 2002; Zhaxybayeva et al., 2006; Andam et al., 2010; Papke and Gogarten, 2012; Williams et al., 2012). Interestingly, for trees where the Halobacteria were polyphyletic, the organisms interrupting the clade were Bacteria for two out of the four polyphyletic intein alleles. The sample size restricts building strong claims about HGT between the Halobacteria and the Bacteria. However, this claim is supported by previous evidence of gene exchange between the Bacteria and the Halobacteria (Ng et al., 2000; Khomyakova et al., 2011).

The tight clustering of halobacterial intein sequences and short branches between closely related strains indicate that in the majority cases inteins are inherited vertically or are transferred between closely related strains, and that successful invasion across large genetic distances is rare. Thus, intein alleles that are found in many different genera have been active for many generations, enabling invasion of many lineages, and accumulating examples of rare invasion events such as those that cross domain boundaries. Conversely, a lack of taxonomic diversity cannot be interpreted as a recent invasion as sampling limitations could be responsible for the paucity of samples in that intein allele. While many factors influence the success of intein transfer between divergent organisms, phylogenetic diversity of the organisms invaded by a particular intein allele also is a reflection of the time the intein allele has been present in a linage. Furthermore, a high density of intein sequences in a particular domain or group of genera can be used to determine the most likely reservoir for the circulating intein allele. A stacked column chart was used to quantify the representation of each of the genera in each of the intein alleles (Figure 4). Five intein alleles, cdc21b, pol-IIa, polBb, cdc21a, and rfc-d, show polarity in intein density favoring the Halobacteria (specifically Halorubrum) as the reservoir for the intein population. This is not surprising as the data indicate that the majority of intein transfer in the Halobacteria is within the class. Additionally, the diversity in five of the intein alleles, helicase-b, cdc21a, gyrB, rir1-b, and udp, suggests these intein populations may be more ancient than the others in this study as they have had time to accumulate rare, long distance transfers such that the diversity within them spans both class and domain boundaries. Interestingly, the helicase-b intein was only recently discovered in this study, though the diversity in the allele gives the impression that this intein has been around for a long time.

TRANSFER OF INTEINS BETWEEN HALOBACTERIAL AND NON-HALOBACTERIAL LINEAGES

Not all inteins are transferred equally; the efficiency of intein invasion is affected largely by the state of the intein. The HEN domain in canonical inteins is required to induce a double strand break and the subsequent homologous repair that results in invasion (Pietrokovski, 2001). Thus, mini-inteins that have lost a functioning HEN domain are mainly transferred vertically (they may be transferred horizontally together with the host gene). If an intein containing allele has been fixed in a population, either a precise deletion of the mini intein encoding DNA could remove the intein from the population or homologous replacement by an intein-free allele transferred from outside the population. Thus, mini-inteins are maintained through strong purifying selection, because any mutation that decreases the self-splicing activity decreases the availability of the host protein (Barzel et al., 2011). The intein states were determined to infer patterns of homing in the Halobacteria. The size of inteins in each allele, along with the position of gaps in the alignment relative to the HEN domain were used as a heuristic for assigning mini-intein status. In most cases there was a clear separation in the distribution of intein lengths (at least 100 amino acids difference in length). The size of more populated intein alleles within the three genera of the Halobacteria with the largest number of available genomes, Haloarcula, Haloferax, and Halorubrum, were recorded in a matrix of intein alleles (Figure 5). Many intein alleles show

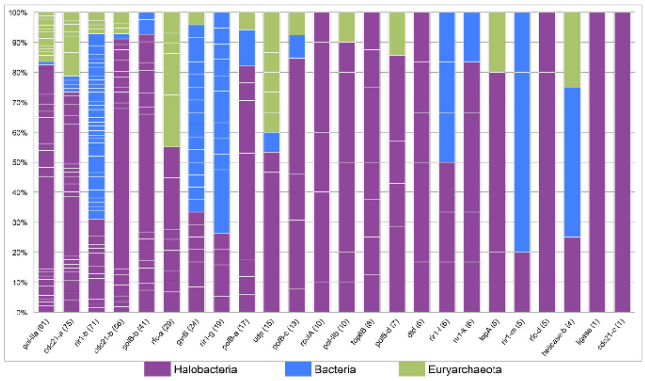


FIGURE 4 | Phylogenetic diversity in halobacterial intein alleles. A stacked column graph depicts the representation of the Halobacteria (in purple), the Bacteria (in blue), and other Euryarchaeota (in green). Intein alleles are ordered by the number of intein sequences recovered for each allele, which is reported in parenthesis after the intein allele name on the x-axis. The number of genera for each intein allele is indicated by the number of breaks in the column (white lines) and the height of each of the fragments that make up a column indicate the proportion of sequences in that allele found in a particular genus.

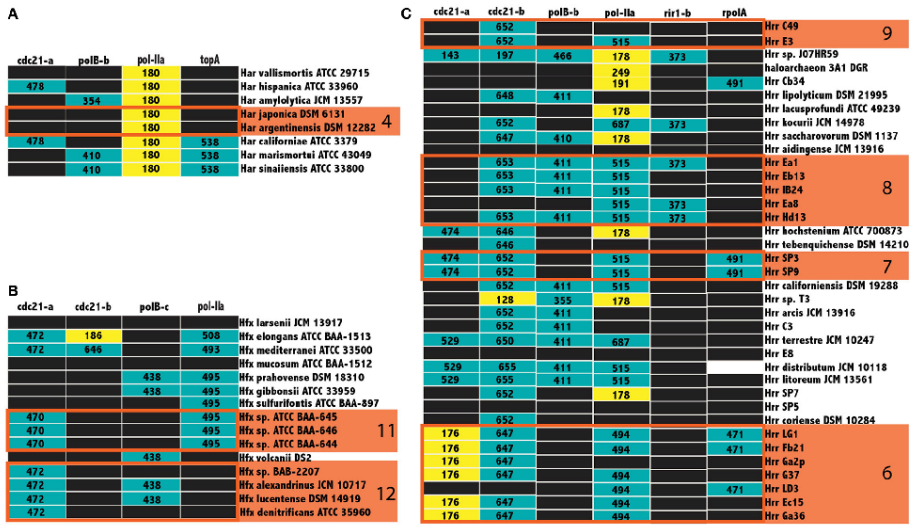


FIGURE 5 | Intein size distributions in the *Haloferax*, *Haloarcula*, and *Halorubrum*. The size of inteins in the *Haloarcula* **(A)**, *Haloferax* **(B)**, and *Halorubrum* **(C)** are indicated in the column corresponding to the intein allele. Mini-inteins are colored yellow, large inteins are colored teal, black boxes indicate no intein, and white boxes indicate missing data, clusters from **Figure 3** are indicated by numbered orange boxes. The *cdc21*-a, and b sequences for *Halorubrum* sp. J07HR59, though smaller than the rest, cannot be considered mini-inteins, as the intein sequences in these positions are not complete.

a considerable size variation. This variability can be attributed to the accumulation of insertions and deletions in various lineages over time, which in some lineages leads to loss of the HEN domain. Notably, there is no variability in the size of intein sequences shared by the clusters recovered in the Bayesian analysis (orange boxes Figure 5) reinforcing the claim of ongoing gene exchange in these clusters.

Invasion from outside the Halobacteria is one explanation for the polyphyletic topology observed in some halobacterial intein alleles. To determine when these homing events could have occurred, the state of each intein was determined and mapped onto polyphyletic intein allele trees: the results of that analysis are summarized in Table 3, with mini-inteins indicated with a star (*), and inteins that group within the Halobacteria indicated by a tilde (~) next to the name of the organism. Many of the intein sequences (5 out of 11) from taxa outside the Halobacteria that interrupt the clade are large-inteins, indicating that interactions between these taxa and the Halobacteria, though rare are ongoing (Table 3). Though the assignment of direction of transfers is extremely preliminary as limited sampling can affect the assignment of direction of transfer, there are some cases with an overwhelming signal where the majority of sequences originate from the Halobacteria, or the Bacteria in the case of rir1-m. The mixture of mini and large inteins represented in all of the intein alleles imply most of these inteins are active in the Halobacteria, and notably involve a wide distribution of taxonomic exchange partners.

DISCUSSION

The importance of HGT throughout the tree of life demands the development of a system to monitor gene-flow within and between populations. This research provides fundamental evidence that mobile elements such as inteins can be used to uncover gene flow networks. Inteins have a unique combination of traits that make them ideal tools to study evolution in microbial populations. They have a naturally wide phylogenetic distribution, enabling detection of HGT between distantly related taxa. This is demonstrated in this work by the intein trees where the Halobacteria were polyphyletic (pol-II a, polB-a, polB-b, and cdc21b) indicating intein transfer between the Halobacteria and the taxa that interrupt them, as well as by data from other studies where intein transfer has been detected across phyla and domains (Butler et al., 2006; Swithers et al., 2013). Inteins also have a high substitution rate relative to their extein hosts, and a propensity for accumulating insertions and deletions, which makes detection of transfers between close relatives (generally a difficult task) possible; for example, transfer within the Halorubrum clusters shown in Figure 3. Inteins can be associated with a HEN domain. If they are, they possess the ability to invade intein-free alleles following transfer; if they are not, they rely mainly on vertical inheritance together with the host gene, and the occasional transfer of the host gene. One intein allele, pol-II a, is widely distributed in the Halobacteria and there are many examples of mini-intein sequences in this allele. These data suggest that invasion of this allele occurred early in the evolution of the Halobacteria, and that the intein may have been lost in some lineages, but retained as a mini intein in most of the genomes surveyed here. This could also be true for the cdc21-a intein; however, the distribution is not as diverse, and considerably fewer mini-inteins were detected. This is more suggestive of an intein that has been active in the Halobacteria for a long period of time, with the different intein states (empty target site, target site invaded by an intein with active HEN, target site occupied by an intein without functioning HEN; Yahara et al., 2009; Barzel et al., 2011) existing and co-existing in different halobacterial lineages.

The genomes analyzed in this work were cultured from salty water and soil samples around the world. The diverse background of the genomes may contribute to the spotty distribution of intein alleles (Figure 1). However, genomes isolated from the same location show variation as well (Figure 3) (Fullmer et al., 2014), reinforcing the notion that inteins are currently actively propagating in and being eliminated from halobacterial populations. Additionally, previous data have shown recombination occurs at a higher rate than mutation within the Halobacteria, and very little linkage between genes is detected in these genomes (Papke et al., 2004, 2007). These observations indicated gene flow as an important method for niche adaptation in these organisms. In Deep Lake, Antarctica the freezing temperatures limit the rate of replication to approximately 6 times per year and evolution in the halobacterial populations there mainly occurs through gene flow (Demaere et al., 2013). Recent whole genome comparisons revealed frequent gene transfer followed by homologous replacement of the transferred gene within the Halobacteria, hampering attempts to resolve the phylogeny within this group (Williams et al., 2012). Gene flow and recombination between populations and species make it difficult to resolve the species phylogeny among the different genera of Halobacteria (Papke et al., 2004). The use of gene concatenation in building reference trees, as exemplified by the ribosomal protein reference tree used in this work, has been pivotal in determining a branching order for the major clades of organisms, such as the Halobacteria, that participate in a large amount of recombination with close relatives. However, because genetic transfer and homologous recombination occur frequently between close relatives, the resulting phylogeny reflects both, shared ancestry and frequency of gene transfer. Therefore, determining the network of gene flow that overlays the vertical signal is important to the understanding of the evolution of these organisms. Inteins cannot penetrate the cell wall, and thus capitalize on existing gene flow in populations to efficiently invade when the opportunity presents itself. This trait can be exploited to keep track of successful homing events revealed by sequence similarity of inteins in distinct strains.

Halorubrum was the only genus in this study that had a large enough sample size to begin to uncover a signal reflecting population structure. Many of the Halorubrum genomes in this study were isolated from the same location, and this collection of genomes showed a clear signal for a structured population. Sixteen genomes from Aran-Bidgol were separated into four well-supported clusters. Three of the four clusters have branching orders identical to those in the reference tree, and the support values for those clusters could be attributed to both transfer within the group and a background phylogenetic signal or ancestral inheritance of similar intein alleles. However, only cluster 7 in the Halorubrum shares all intein alleles between all members of

Table 3 | Protein sequence identifiers for intein sequences.

Intein allele	Species name	Accession number	Phylum
cdc21-a	*Archaeoglobus profundus* DSM 5631	YP_004340760.1	Euryarchaeota
	[#] *Archaeoglobus veneficus* SNP6	YP_003400528.1	Euryarchaeota
	[#] *Candidatus* Methanomassiliicoccus intestinalis Issoire Mx1	YP_008072558.1	Euryarchaeota
	[#] *Croccosphaera watsonii*	WP_021836378.1	Cyanobacteria
	[#] *Ferroglobus placidus* DSM 10642	YP_003435419.1	Euryarchaeota
	[#] *Halarchaeum acidiphilum*	WP_020220725.1	Halobacteria
	[#] *Lamprocystis purpurea*	WP_020504136.1	Gammaproteobacteria
	[#] *Methanomassiliicoccus luminyensis*	WP_019178416.1	Euryarchaeota
	[#] *Methanothermococcus okinawensis* IH1	YP_004576471.1	Euryarchaeota
	Nocardia asteroides NBRC 15531	GAD83132.1	Actinobacteria
	Nocardiopsis potens	WP_020380316.1	Actinobacteria
	Pyrococcus abyssi GE5	NP_127115.1	Euryarchaeota
	[#] *Pyrococcus furiosus* DSM 3638	NP_578211.1	Euryarchaeota
	[#] *Pyrococcus horikoshii* OT3	NP_142122.1	Euryarchaeota
	[#] *Pyrococcus* sp. NA2	YP_004424138.1	Euryarchaeota
	Thermococcus litoralis DSM 5473	YP_0084297171	Euryarchaeota
	[#] *Thermococcus onnurineus* NA1	YP_002306424.1	Euryarchaeota
	[#] *Thermococcus sibiricus* MM 739	YP_002994932.1	Euryarchaeota
	[#] *Thermococcus* sp. AM4	YP_002582218.1	Euryarchaeota
	[#] *Thermococcus* sp. CL1	YP_006424652.1	Euryarchaeota
	[#] *Thermococcus zilligii*	WP_0104791211	Euryarchaeota
	Halorubrum sp. SP3	KJ_8656871	Halobacteria
	Halorubrum sp. SP9	KJ_865689.1	Halobacteria
cdc21-b	[#] *Cyanothece* sp. PCC 7822	YP_0038878971	Cyanobacteria
	Halarchaeum acidiphilum	WP_020220725.1	Halobacteria
	[#] *Candidatus* Methanomassiliicoccus intestinalis Issoire-Mx1	YP_008072558.1	Euryarchaeota
	[#] *Methanomassiliicoccus luminyensis*	WP_019178416.1	Euryarchaeota
	~ *Thermococcus barophilus*	YP_004070279.1	Euryarchaeota
	Halorubrum sp. SP3	KJ_8656871	Halobacteria
	Halorubrum sp. SP7	KJ_865688.1	Halobacteria
	Halorubrum sp. SP9	KJ_865689.1	Halobacteria
polB-d	*Archaeoglobus profundus* DSM 5631	YP_003400528.1	Euryarchaeota
polB-a	~ *Salinibacter ruber* M8	YP_003572085.1	Bacteroidetes
	~ *Salinibacter ruber* DSM 13885	YP_446104.1	Bacteroidetes
	~ *Halarchaeum acidiphilum*	WP_020678478.1	Halobacteria
	~ *Methanoculleus bourgensis*	YP_006544623.1	Euryarchaeota
polB-b	*Halosimplex carlsbadense*	WP_006885382.1	Halobacteria
	[#] ~ *Salinibacter ruber* M8	YP_003572085.1	Bacteroidetes
	[#] ~ *Salinibacter ruber* DSM 13885	YP_446104.1	Bacteroidetes
	[#] ~ *Halanaerobium saccharolyticum*	WP_0054890971	Firmicutes
	Halarchaeum acidiphilum	WP_020678478.1	Halobacteria
polB-c	[#] ~ *Thermus scotoductus*	YP_004202875.1	Deinococcus-Thermus
	[#] ~ *Methanotorris igneus* Kol 5	YP_004483799.1	Euryarchaeota
	Halorubrum sp. SP7	KJ_865686.1	Halobacteria
*pol-II*a	*Archaeoglobus veneficus* SNP6	YP_004341738.1	Euryarchaeota
	Halosimplex carlsbadense	WP_006882195.1	Halobacteria
	[#] *Methanocaldococcus infernus* ME	YP_0036169471	Euryarchaeota
	Methanococcus aeolicus	ABU41683.1	Euryarchaeota

(Continued)

Table 3 | Continued

Intein allele	Species name	Accession number	Phylum
	[#] *Methanoculleus bourgensis* MS2	YP_006544019.1	Euryarchaeota
	[#] *Methanoculleus marisnigri* JR-1	YP_001048029.1	Euryarchaeota
	Methanofollis liminatans	WP_004037227.1	Euryarchaeota
	[#] *Methanolinea tarda*	WP_007314808.1	Euryarchaeota
	[#] *Methanoplanus limicola*	WP_004076782.1	Euryarchaeota
	[#] *Methanoplanus petrolearius* DSM 11571	YP_003893638.1	Euryarchaeota
	Methanoregula boonei 6A8	YP_001403293.1	Euryarchaeota
	~ *Methanoregula formicica* SMSP	YP_007242862.1	Euryarchaeota
	Methanosphaerula palustris E1-9c	YP_002467270.1	Euryarchaeota
	[#] *Metahnospirillum hungatei* JF-1	YP_503855.1	Euryarchaeota
	[#] *Pyrococcus horikoshii* OT3	NP_142130.1	Euryarchaeota
	[#] *Thermococcus gammatolerans* EJ3	YP_002958492.1	Euryarchaeota
	[#] *Thermococcus sibiricus* MM 739	YP_002994988.1	Euryarchaeota
	uncultured haloarchaeon	ABQ75865.1	Halobacteria
	Halorubrum sp. SP3	KJ_865692.1	Halobacteria
	Halorubrum sp. SP7	KJ_865690.1	Halobacteria
	Halorubrum sp. SP9	KJ_564691.1	Halobacteria
pol-IIb	*Halosimplex carlsbadense*	WP_006882195.1	Halobacteria
	[#] *Pyrococcus abyssi* GE5	YP_004624494.1	Euryarchaeota
	uncultured haloarchaeon	ABQ75865.1	Halobacteria
gyrB	*Allochromatium vinosum* DSM 180	YP_003443943.1	Gammaproteobacteria
	Anabaena sp. 90	YP_006997726	Cyanobacteria
	[#] *Anabaena* sp. PCC 7108	WP_016950132.1	Cyanobacteria
	Bacillus subtilis BEST7613	BAM51471.1	Firmicutes
	Calothrix sp. PCC 7103	WP_019489451.1	Cyanobacteria
	Coleofasciculus chthonoplastes	WP_006099284.1	Cyanobacteria
	[#] *Cylindrospermopsis reciborskii*	WP_006276716.1	Cyanobacteria
	[#] *Dactylococcopsis slaina* PCC 8305	YP_007173052.1	Cyanobacteria
	Halarchaeum acidiphilum	WP_021780646.1	Halobacteria
	Methanomassiliicoccus luminyensis	WP_019178436.1	Euryarchaeota
	Microcystis aeruginosa	WP_002774451.1	Cyanobacteria
	Moorea producens	WP_008190351.1	Cyanobacteria
	Oscillatoria sp. PCC 10802	WP_017715151.1	Cyanobacteria
	Pleurocapsa sp. PCC 7319	WP_019509077.1	Cyanobacteria
	Prochlorothrix hollandica	WP_017710941.1	Cyanobacteria
	Raphidiopsis brookii	WP_009342634.1	Cyanobacteria
	Rivularia sp. PCC 7116	YP_007054134.1	Cyanobacteria
	Saccharothrix espanaensis DSM 44229	YP_007037469.1	Actinobacteria
	Synechocystis sp. PCC 6803	NP_441040.1	Cyanobacteria
	Trichodesium erythraeum IMS101	YP_723459.1	Cyanobacteria
	uncultured bacterium	EKD46222.1	
helicase-b	[#] *Bacillus amyloliqufaciens* TA208	YP_005540906.1	Firmicutes
	[#] *Bacillus subtilis*	WP_017696872.1	Firmicutes
	Nanoarchaeota archaeon SCGC AAA011-L22	WP_018204386.1	
rfc-a	*Methanocaldococcus jannaschii* DSM 2661	NP_248426.1	Euryarchaeota
	Methanocaldococcus sp. FS406	YP_003458055.1	Euryarchaeota
	Methanothermococcus okinawensis IH1	YP_004576337.1	Euryarchaeota
	[#] *Methanotorris formicicus*	WP_007044297.1	Euryarchaeota
	[#] *Pyrococcus abyssi* GE5	NP_125803.1	Euryarchaeota

(Continued)

Table 3 | Continued

Intein allele	Species name	Accession number	Phylum
	[#] *Pyrococcus furiosus* DSM 3638	NP_577822.1	Euryarchaeota
	[#] *Pyrococcus horikoshii* OT3	NP_142122.1	Euryarchaeota
	[#] *Pyrococcus* sp. ST04	YP_006353924.1	Euryarchaeota
	[#] *Thermococcus kodakorensis* KOD1	YP_184631.1	Euryarchaeota
	[#] *Thermococcus litoralis* DSM 5473	YP_0084288971	Euryarchaeota
	Thermococcus sp. 4557	YP_004763272.1	Euryarchaeota
	Thermococcus sp. AM4	YP_002582171.1	Euryarchaeota
	[#] *Thermococcus* sp. CL1	YP_006425306.1	Euryarchaeota
*rpoI*A	*Halorubrum* sp. SP3	KJ_865684.1	Halobacteria
	Halorubrum sp. SP9	KJ_865685.1	Halobacteria
rir1-l	*Chloroherpeton thalassium* ATCC 35110	YP_001995975.1	Chlorobi
	Tepidanaerobacter acetatoxydans Re1	YP_007273179.1	Firmicutes
	uncultured Chloroflexi bacterium	BAL53207.1	Chloroflexi
rir1-k	*Deinococcus peraridilitoris* DSM 19664	YP_007181218.1	Deinococcus-Thermus
rir1-b	*Acidovorax avenae* subsp. avenae ATCC 19860	YP_004233126.1	Betaproteobacteria
	Acidovorax sp. CF316	WP_007856012.1	Betaproteobacteria
	Acidovorax sp. NO-1	WP_008903130.1	Betaproteobacteria
	Actinomadura atramentaria	WP_019631066.1	Actinobacteria
	Alicyclobacillus pohliae	WP_018131875.1	Firmicutes
	Aminomonas paucivorans	WP_006300529.1	Synergistetes
	Ammonifex degensii KC4	WP_006300529.1	Firmicutes
	Arhodomonas aquaeolei	WP_018718131.1	Gammaproteobacteria
	Bacillus licheniformis	WP_016885361.1	Firmicutes
	Bacillus subtilis	WP_017697104.1	Firmicutes
	Calothrix sp. PCC 6303	YP_007136749.1	Cyanobacteria
	Candidatus Chloracidobacterium thermophilum B	YP_004863563.1	Acidobacteria
	Candidatus Desulforudis audaxviator MP104C	YP_001717412.1	Firmicutes
	Clostridiaceae bacterium L21-TH-D2	WP_006314960.1	Firmicutes
	Deinococcus radiodurans R1	NP_296095.1	Deinococcus-Thermus
	Delftia acidovorans	WP_016451949.1	Betaproteobacteria
	Delftia sp. Cs1-4	YP_004490724.1	Betaproteobacteria
	Desulfitobacterium hafniense	WP_005810476.1	Firmicutes
	Desulfovibrio magneticus RS-1	YP_002955841.1	Deltaproteobacteria
	Desulfovibrio sp. U5L	WP_009106508.1	Deltaproteobacteria
	Ferroplasma acidarmanus fer1	YP_008141532.1	Euryarchaeota
	Ferroplasma sp. Type II	WP_021787573.1	Euryarchaeota
	Halomonas anticariensis	WP_016418429.1	Gammaproteobacteria
	Halomonas jeotgali	WP_017429019.1	Gammaproteobacteria
	Halomonas smyrnensis	WP_016854101.1	Gammaproteobacteria
	Mahella australiensis 50-1 BON	YP_004462974.1	Firmicutes
	Marinobacter lipolyticus	WP_018405479.1	Gammaproteobacteria
	Methanofollis liminatans	WP_004040239.1	Euryarchaeota
	Methylobacter marinus	WP_020160338.1	Gammaproteobacteria
	Methylococcus capsulatus	WP_017366201.1	Gammaproteobacteria
	Methylomicrobium buryatense	WP_017841702.1	Gammaproteobacteria
	nanoarchaeote Nst1	WP_0045780171	
	Nocardiopsis halotolerans	WP_0175723471	Actinobacteria
	Polaromonas sp. JS666	CAJ571771	Cyanobacteria
	Pseudanabaena sp. PCC 6802	WP_019499030.1	Cyanobacteria

(Continued)

Table 3 | Continued

Intein allele	Species name	Accession number	Phylum
	Pseudanabaena sp. PCC 7367	YP_007101092.1	Cyanobacteria
	Rhodanobacter fulvus	WP_007082010.1	Gammaproteobacteria
	Rhodanobacter sp. 2APBS1	YP_007588821.1	Gammaproteobacteria
	Rhodanobacter thiooxydans	WP_008437232.1	Gammaproteobacteria
	Rhodothermus marinus SG0.5JP17-172	YP_004824118.1	Bacteroidetes
	Staphylococcus aureus	WP_016187732.1	Firmicutes
	Synechococcus elongatus PCC 6301	CAJ57178.1	Cyanobacteria
	Synechococcus elongatus PCC 7942	YP_400626.1	Cyanobacteria
	Synechococcus sp. PCC 6312	YP_007060778.1	Cyanobacteria
	Thermoanaerobacterium saccharolyticum JW/SL-YS485	YP_006391581.1	Firmicutes
	Thermoanaerobacterium thermosaccharolyticum DSM 571	YP_003851043.1	Firmicutes
	Thermobrachium celere	WP_018663796.1	Firmicutes
	Thermococcus kodakarensis KOD1	YP_184312.1	Euryarchaeota
	Thermodesulfatator indicus DSM 15286	YP_004625205.1	Thermodesulfobacteria
	Thermovirga lienii DSM 17291	YP_004932130.1	Deinococcus-Thermus
	Thermus igniterrae	WP_018110436.1	Deinococcus-Thermus
	Thermus thermophilus HB8	CAJ57170.1	Deinococcus-Thermus
	Thioalkalivibrio sp. ALE11	WP_019570879.1	Gammaproteobacteria
	Thioalkalivibrio sp. ALE30	WP_018881426.1	Gammaproteobacteria
	Thioalkalivibrio sp. HL-Eb18	WP_017926201.1	Gammaproteobacteria
	Thioalkalivibrio sp. K90mix	YP_003459507.1	Gammaproteobacteria
	uncultured bacterium	EKE25755.1	
	Xanthomonas sp. SHU199	WP_017907463.1	Gammaproteobacteria
	Xanthomonas sp. SHU308	WP_017915139.1	Gammaproteobacteria
	zeta proteobacterium SCGC AB-604-B04	WP_018280466.1	Zetaproteobacteria
rir1-g	*Chloroherpeton thalassium* ATCC 35110	YP_001995975.1	Chlorobi
	Deinococcus aquatillis	WP_019011777.1	Deinococcus-Thermus
	Halothece sp. PCC 7418	YP_007166732.1	Cyanobacteria
	Klebsiella pneumoniae	WP_021313783.1	Gammaproteobacteria
	Nocardiopsis dassonvillei subsp. Dassonvillei DSM 43111	YP_003681238.1	Actinobacteria
	Nocardiopsis sp. CNS639	WP_019609645.1	Actinobacteria
	Rhodothermus marinus SG0.5JP17-172	YP_004826277.1	Bacteroidetes
	Tepidanaerobacter acetatoxydans Re1	YP_007273179.1	Firmicutes
	Thermomonospora curvata DSM 43183	YP_003299200.1	Actinobacteria
	Thermus thermophilus HB27	YP_005899.1	Deinococcus-Thermus
	Thermus thermophilus HB8	CAJ57173.1	Deinococcus-Thermus
	Thermus thermophilus JL-18	YP_006059430.1	Deinococcus-Thermus
	Thermus thermophilus SG0.5JP17-16	YP_005639869.1	Deinococcus-Thermus
	Trichodesium erythraeum IMS101	YP_720358.1	Cyanobacteria
	uncultured Chloroflexi bacterium	BAL53207.1	Chloroflexi
rir1-m	*Thermus aquaticus*	WP_003044118.1	Deinococcus-Thermus
	Thermus thermophilus HB-8	CAJ57173.1	Deinococcus-Thermus
	Thermus thermophilus SG0.5JP17-16	YP_005639869.1	Deinococcus-Thermus
	uncultured Chloroflexi bacterium	BAL53207.1	Chloroflexi
udp	*Fervidibacteria bacterium* JGI 0000001-G10	WP_020250137.1	
	Dictyglomus thermophilum H-6-12	YP_002250310.1	Dictyglomi
	Methanocaldococcus jannaschii DSM 2661	NP_248048.1	Euryarchaeota
	Methanocaldococcus vulcanis M7	YP_003246412.1	Euryarchaeota
	Methanococcus aeolicus Nankai-3	YP_001324612.1	Euryarchaeota
	Methanothermococcus okinawensis IH1	YP_004575831.1	Euryarchaeota

(Continued)

Table 3 | Continued

Intein allele	Species name	Accession number	Phylum
	Methanotorris igneus Kol 5	WP_007044255.1	Euryarchaeota
	Thermococcus gammatolerans EJ3	YP_002960518.1	Euryarchaeota
topA	*Methanotorris igneus* Kol 5	WP_007044255.1	Euryarchaeota
top6B	*Halarchaeum acidiphilum*	WP_021780130.1	Halobacterium

**Indicates the intein detected is a mini-intein.*

~ Indicates taxa that grouped within the halobacterial intein sequences.

the cluster while the other clusters all contain intein alleles that are unique to certain members of the cluster, suggesting ongoing transfer of these inteins within the population. Additionally, three out of the twelve total clusters demonstrate unique branching orders compared to the reference tree, though only five of the clusters reflected in the reference tree have identical intein profiles. The lack of fixation for the intein alleles in the majority of clusters (seven out of twelve) indicates that a signal due to vertical inheritance may aid the formation of the clusters, but that HGT and its bias is the driving force for intein distribution. This analysis demonstrates the utility of intein sequences in distinguishing a population structure amongst genomes isolated from the same location, as demonstrated with the genomes isolated from Aran-Bidgol. These relationships are made evident through analyzing all of the signals from each of the intein alleles represented in the strains, and thus represent a collapsed view of the major gene sharing networks that have shaped the intein profiles of these strains over time. The collapsed networks indicate a higher rate of recombination within compared to between species and groups, a finding similar to the sexual outcrossing in fungal populations where inteins also thrive, as the semi-sexual lifestyle promotes intein homing (Giraldo-Perez and Goddard, 2013).

It is tempting to speculate that strains that harbor an abundance of intein alleles partake in more gene transfer than their counterparts without as many inteins; however, these two phenomena should not be expected to have a strict correlation as HGT between strains that possess only one intein each cannot produce hybrids with more than two inteins each. The number of inteins present in a group of different strains and species may be more reflective of transfers with divergent organisms than within-group transfer frequency.

The presented research demonstrates the utility of intein sequences to follow gene flow within and between populations. Improved reliability to assess the presence and activity of the HEN domain intein will provide a better distinction between vertical and horizontal inheritance of inteins. The overall utility of inteins improves as new intein alleles and new host proteins are reported, increasing the distribution of samples and improving statistical robustness of studies like the one done here. Prior to this work, nine proteins had been reported to contain inteins in the Halobacteria. This work established seven new intein alleles in the Halobacteria, including two proteins not previously reported to contain inteins. The presence of inteins is especially useful in populations where high rates of recombination and widely distributed populations may facilitate the maintenance of intein sequences over long periods of time (Gogarten and Hilario, 2006) and provide a means for distinguishing closely related partners involved in genetic transfers. The phylogenetic distribution of intein alleles, combined with the changing state within intein alleles, and the rapid substitution rate of inteins relative to the extein host sequences (Swithers et al., 2013) will provide a valuable tool to infer gene flow dynamics in and between sampled populations.

AUTHOR CONTRIBUTIONS

Johann Peter Gogarten and Shannon M. Soucy participated in the design of this study and helped to draft the manuscript. Shannon M. Soucy performed the research and all authors contributed to data analysis. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

The UConn Bioinformatics Facility provided computing resources for the analyses reported in this manuscript. The Halorubrum genomes provided by the Papke lab were sequenced in house by Andrea Makkay and Ryan Wheeler. We would like to thank them for their hard work, as well as acknowledge Dr. Elina Roine and Dennis Bamford (Helsinki University), and Dr. Antonio Ventosa (University of Sevilla) for supplying the sequenced strains. We would also like to recognize labs sequencing genomes and making them available in data repositories such as those hosted by the National Center for Biotechnology Information. This work was supported by the National Science Foundation Grant (DEB 0830024 and DEB0919290) and NASA Astrobiology: Exobiology and Evolutionary Biology Grants (NNX12AD70G and NNX13AI03G).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2014.00299/abstract

REFERENCES

Altschul, S. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Andam, C. P., Williams, D., and Gogarten, J. P. (2010). Biased gene transfer mimics patterns created through shared ancestry. Proc. Natl. Acad. Sci. U.S.A. 107, 10679–10684. doi: 10.1073/pnas.1001418107

Atanasova, N. S., Roine, E., Oren, A., Bamford, D. H., and Oksanen, H. M. (2012). Global network of specific virus-host interactions in hypersaline environments. Environ. Microbiol. 14, 426–440. doi: 10.1111/j.1462-2920.2011.02603.x

Barzel, A., Obolski, U., Gogarten, J. P., Kupiec, M., and Hadany, L. (2011). Home and away- the evolutionary dynamics of homing endonucleases. BMC Evol. Biol. 11:324. doi: 10.1186/1471-2148-11-324

Butler, M. I., Gray, J., Goodwin, T. J., and Poulter, R. T. (2006). The distribution and evolutionary history of the PRP8 intein. BMC Evol. Biol. 6:42. doi: 10.1186/1471-2148-6-42

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27, 1164–1165. doi: 10.1093/bioinformatics/btr088

Demaere, M. Z., Williams, T. J., Allen, M. A., Brown, M. V., Gibson, J. A. E., Rich, J., et al. (2013). High level of intergenera gene exchange shapes the evolution of haloarchaea in an isolated Antarctic lake. Proc. Natl. Acad. Sci. U.S.A. 110, 16939–16944. doi: 10.1073/pnas.1307090110

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. doi: 10.1093/nar/gkh340

Fullmer, M. S., Soucy, S. M., Swithers, K. S., Makkay, A. M., Wheeler, R., Ventosa, A., et al. (2014). Population and genomic analysis of the genus Halorubrum. Front. Microbiol. 5:140. doi: 10.3389/fmicb.2014.00140

Gimble, F. S., and Thorner, J. (1992). Homing of a DNA endonuclease gene by meiotic gene conversion in Saccharomyces cerevisiae. Nature 357, 301–306. doi: 10.1038/357301a0

Giraldo-Perez, P., and Goddard, M. R. (2013). A parasitic selfish gene that affects host promiscuity. Proc. Biol. Sci. 280:20131875. doi: 10.1098/rspb.2013.1875

Goddard, M. R., and Burt, A. (1999). Recurrent invasion and extinction of a selfish gene. Proc. Natl. Acad. Sci. 96, 13880–13885. doi: 10.1073/pnas.96.24.13880

Gogarten, J. P., and Hilario, E. (2006). Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements. BMC Evol. Biol. 6:94. doi: 10.1186/1471-2148-6-94

Gogarten, J. P., Senejani, A. G., Zhaxybayeva, O., Olendzenski, L., and Hilario, E. (2002). Inteins: structure, function, and evolution. Annu. Rev. Microbiol. 56, 263–287. doi: 10.1146/annurev.micro.56.012302.160741

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27, 221–224. doi: 10.1093/molbev/msp259

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321. doi: 10.1093/sysbio/syq010

Hirata, R., Ohsumk, Y., Nakano, A., Kawasaki, H., Suzuki, K., and Anraku, Y. (1990). Molecular structure of a gene, VMA1, encoding the catalytic subunit of H(+)-translocating adenosine triphosphatase from vacuolar membranes of Saccharomyces cerevisiae. J. Biol. Chem. 265, 6726–6733.

Kane, P. M., Yamashiro, C. T., Wolczyk, D. F., Neff, N., Goebl, M., and Stevens, T. H. (1990). Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H(+)-adenosine triphosphatase. Science 250, 651–657. doi: 10.1126/science.2146742

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780. doi: 10.1093/molbev/mst010

Khomyakova, M., Bükmez, Ö., Thomas, L. K., Erb, T. J., and Berg, I. A. (2011). A methylaspartate cycle in haloarchaea. Science 331, 334–337. doi: 10.1126/science.1196544

Lang, A. S., Zhaxybayeva, O., and Beatty, J. T. (2012). Gene transfer agents: phage-like elements of genetic exchange. Nat. Rev. Microbiol. 10, 472–482. doi: 10.1038/nrmicro2802

Liu, K., Warnow, T. J., Holder, M. T., Nelesen, S. M., Yu, J., Stamatakis, A. P., et al. (2012). SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. Syst. Biol. 61, 90–106. doi: 10.1093/sysbio/syr095

Ng, W. V., Kennedy, S. P., Mahairas, G. G., Berquist, B., Pan, M., Shukla, H. D., et al. (2000). Genome sequence of Halobacterium species NRC-1. Proc. Natl. Acad. Sci. U.S.A. 97, 12176–12181. doi: 10.1073/pnas.190337797

Papke, R. T., and Gogarten, J. P. (2012). Ecology. How bacterial lineages emerge. Science 336, 45–46. doi: 10.1126/science.1219241

Papke, R. T., Koenig, J. E., Rodriguez-Valera, F., and Doolittle, W. F. (2004). Frequent recombination in a saltern population of halorubrum. Science 306, 1928–1929. doi: 10.1126/science.1103289

Papke, R. T., Zhaxybayeva, O., Feil, E. J., Sommerfeld, K., Muise, D., and Doolittle, W. F. (2007). Searching for species in haloarchaea. Proc. Natl. Acad. Sci. U.S.A. 104, 14092–14097. doi: 10.1073/pnas.0706358104

Perler, F. B. (2002). InBase: the Intein Database. Nucleic Acids Res. 30, 383–384. doi: 10.1093/nar/30.1.383

Perler, F. B., Olsen, G. J., and Adam, E. (1997). Compilation and analysis of intein sequences. Nucleic Acids Res. 25, 1087–1093. doi: 10.1093/nar/25.6.1087

Pietrokovski, S. (2001). Intein spread and extinction in evolution. Trends Genet. 17, 465–472. doi: 10.1016/S0168-9525(01)02365-4

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539–542. doi: 10.1093/sysbio/sys029

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539. doi: 10.1038/msb.2011.75

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Swithers, K. S., Senejani, A. G., Fournier, G. P., and Gogarten, J. P. (2009). Conservation of intron and intein insertion sites: implications for life histories of parasitic genetic elements. BMC Evol. Biol. 9:303. doi: 10.1186/1471-2148-9-303

Swithers, K. S., Soucy, S. M., Lasek-Nesselquist, E., Lapierre, P., and Gogarten, J. P. (2013). Distribution and evolution of the mobile vma-1b intein. Mol. Biol. Evol. 30, 2676–2687. doi: 10.1093/molbev/mst164

Williams, D., Gogarten, J. P., and Papke, R. T. (2012). Quantifying homologous replacement of loci between haloarchaeal species. Genome Biol. Evol. 4, 1223–1244. doi: 10.1093/gbe/evs098

Yahara, K., Fukuyo, M., Sasaki, A., and Kobayashi, I. (2009). Evolutionary maintenance of selfish homing endonuclease genes in the absence of horizontal transfer. Proc. Natl. Acad. Sci. U.S.A. 106, 18861–18866. doi: 10.1073/pnas.0908404106

Zhaxybayeva, O., Gogarten, J. P., Charlebois, R. L., Doolittle, W. F., and Papke, R. T. (2006). Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res. 16, 1099–1108. doi: 10.1101/gr.5322306

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 02 January 2014; accepted: 30 May 2014; published online: 26 June 2014.
Citation: Soucy SM, Fullmer MS, Papke RT and Gogarten JP (2014) Inteins as indicators of gene flow in the halobacteria. Front. Microbiol. 5:299. doi: 10.3389/fmicb.2014.00299
This article was submitted to Extreme Microbiology, a section of the journal Frontiers in Microbiology.

Appendix C – Collins et al., 2014

Comparative genomics of Roseobacter clade bacteria isolated from the accessory nidamental gland of Euprymna scolopes

This section features Andrew Collins' manuscript comparing the genomes and genomic complements of alpha-proteobacteria isolates he sequenced from the Hawaiian Bobtail Squid, *Euprymna scalopes*. My contribution to this article was centered around providing a phylogeny which we used to place his isolates as well as whole-genome comparisons with which we could classify the taxa into species. The latter consisted of using jANI in normal order. The former involved adapting/devising a MLSA scheme which fit our particular range of organisms. This boiled down to identifying what MLSA schema had been used in the literature and identifying which genes had good representation in our set of genomes. I participated in the drafting of the relevant areas of the manuscript as well as in the editing of the document.

frontiers in MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 23 February 2015
doi: 10.3389/fmicb.2015.00123

Comparative genomics of *Roseobacter* clade bacteria isolated from the accessory nidamental gland of *Euprymna scolopes*

Andrew J. Collins[1,2], Matthew S. Fullmer[1], Johann P. Gogarten[1,3] and Spencer V. Nyholm[1]*

[1] Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA
[2] Microbiology, The Forsyth Institute, Cambridge, MA, USA
[3] Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA

Edited by:
Shana Goffredi, Occidental College, USA

Reviewed by:
Haiwei Luo, The Chinese University of Hong Kong, China
Wesley Douglas Swingley, Northern Illinois University, USA

***Correspondence:**
Spencer V. Nyholm, Molecular and Cell Biology, University of Connecticut, 91 North Eagleville Road, Unit 3125, Storrs, CT, USA
e-mail: spencer.nyholm@uconn.edu

The accessory nidamental gland (ANG) of the female Hawaiian bobtail squid, *Euprymna scolopes*, houses a consortium of bacteria including members of the *Flavobacteriales*, *Rhizobiales*, and *Verrucomicrobia* but is dominated by members of the *Roseobacter* clade (Rhodobacterales) within the *Alphaproteobacteria*. These bacteria are deposited into the jelly coat of the squid's eggs, however, the function of the ANG and its bacterial symbionts has yet to be elucidated. In order to gain insight into this consortium and its potential role in host reproduction, we cultured 12 Rhodobacterales isolates from ANGs of sexually mature female squid and sequenced their genomes with Illumina sequencing technology. For taxonomic analyses, the ribosomal proteins of 79 genomes representing both roseobacters and non-roseobacters along with a separate MLSA analysis of 33 housekeeping genes from *Roseobacter* organisms placed all 12 isolates from the ANG within two groups of a single *Roseobacter* clade. Average nucelotide identity analysis suggests the ANG isolates represent three genera (*Leisingera*, *Ruegeria*, and *Tateyamaria*) comprised of seven putative species groups. All but one of the isolates contains a predicted Type VI secretion system, which has been shown to be important in secreting signaling and/or effector molecules in host–microbe associations and in bacteria–bacteria interactions. All sequenced genomes also show potential for secondary metabolite production, and are predicted to be involved with the production of acyl homoserine lactones (AHLs) and/or siderophores. An AHL bioassay confirmed AHL production in three tested isolates and from whole ANG homogenates. The dominant symbiont, *Leisingera* sp. ANG1, showed greater viability in iron-limiting conditions compared to other roseobacters, possibly due to higher levels of siderophore production. Future comparisons will try to elucidate novel metabolic pathways of the ANG symbionts to understand their putative role in host development.

Keywords: symbiosis, *Euprymna scolopes*, *Roseobacter* clade, genomics, Cephalopoda, *Alphaproteobacteria*

INTRODUCTION

The *Roseobacter* clade is a pervasive and diverse group of marine *Alphaproteobacteria*. This group is estimated to account for 10% of all marine bacteria, with higher percentages in coastal seawater (Wagner-Döbler and Biebl, 2006). These organisms have usually been investigated from an ecological perspective due to their abundance in seawater. The combined metabolic potential of such a large bacterial population may contribute to both sulfur cycling, primarily through metabolism of dimethylsulfoniopropionate (DMSP), and carbon cycling, as roseobacters oxidize a variety of carbon sources to CO_2 (González et al., 2000).

Many of the characterized *Roseobacter* isolates can be described as free-living, having been isolated from seawater or inert marine surfaces. However, some roseobacters also associate with other organisms, including oysters (Ruiz-Ponte et al., 1998), sponges (Zan et al., 2014), algae (Rao et al., 2007; Case et al., 2011), and cephalopods (Grigioni et al., 2000; Pichon et al., 2005; Collins et al., 2012). Among many squid and cuttlefish, roseobacters have been found associated with the accessory nidamental gland (ANG), part of the female reproductive system and comprised of many epithelium-lined tubules that house dense populations of bacterial symbionts (**Figure 1**, Bloodgood, 1977; Collins et al., 2012). Evidence suggests that these bacteria are embedded in the jelly coat of the squid's eggs that are then deposited in masses on the ocean floor where they resist fouling and degradation over ~3 weeks of development (Barbieri et al., 2001; Collins et al., 2012).

Studies that have investigated the ANG consortium have found members of the *Roseobacter* clade among many cephalopods, including *Doryteuthis pealeii*, *Sepia officinalis*, and *Euprymna scolopes* (Grigioni et al., 2000; Barbieri et al., 2001; Pichon et al., 2005; Collins et al., 2012). In the Hawaiian bobtail squid, *E. scolopes*, roseobacters comprise ~50% of the microbial population according to 16S rDNA surveys, predominantly from the genus *Leisingera* (formerly *Phaeobacter*; Collins et al., 2012). Other members of the consortium include *Flavobacteria* and *Verrucomicrobia*

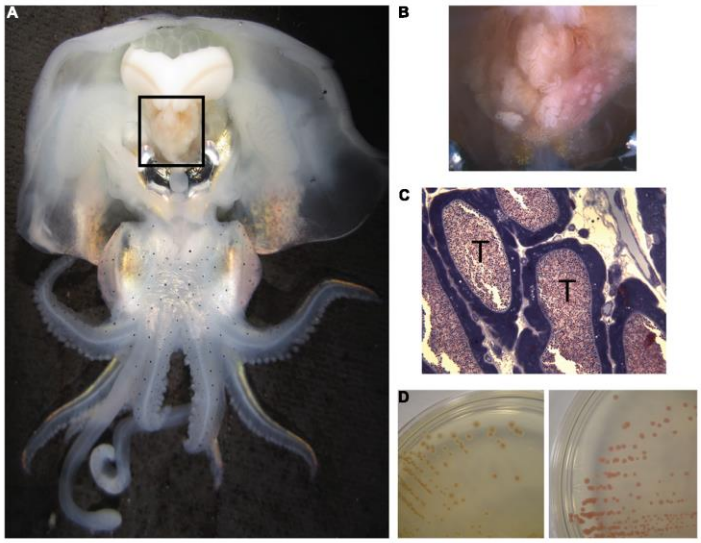


FIGURE 1 | Anatomy of *Euprymna scolopes*, the accessory nidamental gland (ANG) and associated bacteria. (A) The ANG (black box) is located on the ventral side of the squid, posterior to the ink sac and light organ. **(B)** The ANG is comprised of many pigmented convoluted epithelium-lined tubules that may appear white, pink, or brown. **(C)** Microscopy of fixed and stained sections of the ANG reveal dense populations of bacteria within tubules (T). **(D)** Bacteria from the ANG that were isolated on complex media also had a variety of pigments, from brown (*Leisingera* sp.ANG-Vp, left) to red (*Tateyamaria* sp. ANG-S1, right).

and each of these groups are partitioned such that only one taxon dominates any given tubule (Collins et al., 2012).

Roseobacter clade bacteria are known to produce several antimicrobial compounds, including tropodithietic acid (TDA), which has antimicrobial and anti-algal properties (Brinkhoff et al., 2004). Under certain conditions, likely when associated with dying algae, *Phaeobacter inhibens* can also produce anti-algal compounds known as roseobacticides derived from *p*-coumaric acid, a product of lignin degradation (Seyedsayamdost et al., 2011). *Leisingera* sp. Y4I and *Leisingera daeponensis* produce indigoidine, an antimicrobial blue pigment that is synthesized from a unique polyketide/non-ribosomal peptide synthase gene cluster and has been shown to inhibit marine bacteria, including *Vibrio fischeri* (Cude et al., 2012; Dogs et al., 2013).

The function of the ANG and its associated bacterial population remains unknown although protective roles against predation and/or fouling have been suggested (Biggs and Epel, 1991). The distribution of roseobacters among cephalopod ANGs suggests that they have a conserved function in these animals. Furthermore, they must contain traits that allow them to survive in multiple habitats such as seawater, a specialized organ such as the ANG, and within squid egg jelly coats. To shed light on the metabolic capabilities of these bacteria and investigate possible adaptations to living in these different habitats, we examined the genomes of 12 isolates from the ANG of *E. scolopes* and compared them to others from the *Roseobacter* lineage. Here, we describe the genetic content from this select group of roseobacters that exist in conserved symbioses with cephalopods worldwide.

MATERIALS AND METHODS

CULTURING BACTERIA FROM THE ANG

Animals were collected in sand shallows on Oahu, Hawaii and maintained in artificial aquaria as previously described Schleicher and Nyholm (2011). To obtain ANGs, five mature females were anesthetized in Instant Ocean with 2% ethanol. Organs were removed and surface sterilized with 70% ethanol before being homogenized in filter-sterilized squid Ringer's solution (530 mM NaCl, 25 mM $MgCl_2$, 10 mM $CaCl_2$, 20 mM HEPES, pH = 7.5). Tissue homogenate was serially diluted and plated on either salt water tryptone (SWT) or Reasoner's 2A medium (R2A) supplemented with a 70:30 mixture of Instant Ocean and distilled water (Reasoner and Geldreich, 1985; Nyholm et al., 2009). Plates were incubated aerobically at 28°C for 2–7 days. For each animal, colonies with different morphology and/or color were isolated for further analysis.

GENOME SEQUENCING AND ANNOTATION

Genomic DNA was isolated using the MasterPure DNA Extraction kit (Epicentre) from liquid cultures of ANG bacteria grown

overnight at 28°C in either SWT or R2A. DNA was quantified using a Qubit fluorescence assay (Invitrogen). Illumina sequencing libraries were created from 1 ng of genomic DNA using the Nextera XT library kit and the libraries were quantified by a HS DNA Bioanalyzer assay (Agilent). Libraries were sequenced on an Illumina MiSeq sequencer using 2 × 250 bp reads. Draft genomes were assembled using the CLC Genomic Workbench (CLC) using default parameters. For *Leisingera* sp. ANG1 (formerly *Phaeobacter gallaeciensis* ANG1), additional sequencing data was added from a previous sequencing effort using an Illumina mated-pair library (Collins and Nyholm, 2011). Assemblies were annotated using the Rapid Annotation using Subsytem Technology (Aziz et al., 2008, RAST, rast.nmpdr.org) server. To search for Type IV secretion systems (T4SS), the VirB4 protein from *P. inhibens* DSM17395 was used to query the ANG isolate genomes using tblastn. Genomes were also analyzed with Anti-SMASH (Blin et al., 2013, Antibiotic and Secondary Metabolite Analysis Shell, antismash.secondarymetabolites.org) and BAGEL3 (van Heel et al., 2013, BActeriocin Genome mining tool, bagel.molgenrug.nl) for secondary metabolite and bacteriocin biosynthesis gene clusters. Draft genome assemblies have been deposited in DDBJ/EMBL/GenBank under accession numbers AFCF00000000 and JWLC00000000-JWLM00000000. The versions described in this manuscript are AFCF02000000 and JWLC01000000-JWLM01000000.

TAXONOMIC ANALYSIS

A total of 79 genomes were used for analyses in this study. Fifty-seven *Roseobacter* genomes and 10 non-*Roseobacter* genomes were obtained from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/, listed in Supplementary Figure 1). Twelve *Roseobacter* genomes are new to this study, including an improved assembly of the previously published *Leisingera* sp. ANG1 (**Table 1**). To ensure equal gene calling across the genomes, all genomes, including the 67 draft and completed genomes obtained from the NCBI ftp, were re-annotated using the RAST server (Aziz et al., 2008). Assembled contigs were reconstructed from the RAST-generated GenBank files for all genomes using the seqret application of the EMBOSS package (Rice et al., 2000).

An initial survey of the *Roseobacter* clade was made using 51 ribosomal proteins. Queries were obtained from the BioCyc database (Caspi et al., 2010) for *Roseobacter denitrificans* OCh 114, excluding methyltransferases and putative proteins. Unlike many previous studies (Soucy et al., 2014) nucleotide sequences were used to potentially allow finer resolution of relationships. The top hits for each gene were aligned separately using MUSCLE (Edgar, 2004) and evaluated by hand to verify that the sequences were homologs. In-house python scripts created a concatenated alignment from all 51 genes. An optimal model of evolution was determined using the akaike information criterion with correction for small sample size (AICc). The program jModelTest 2.1.4. was used to compute likelihoods from the nucleotide alignment and to perform the AICc (Guindon et al., 2010; Darriba et al., 2012). The best-fitting model reported was GTR + Gamma estimation + Invariable site estimation. A maximum likelihood (ML) phylogeny was generated from the concatenated multi-sequence alignment using PhyML v3.0_360-500M (Guindon et al., 2010). PhyML parameters consisted of GTR model, estimated p-invar, 4 substitution rate categories, estimated gamma distribution, subtree pruning and regrafting enabled with 100 bootstrap replicates. This tree (Supplementary Figure 1) placed all of the new ANG isolates from this study into a single clade, corresponding to three groups (Clades 1, 2, and 4) previously described by Newton et al. (2010). Clade 4's placement sister to clade 2 is discussed in Section "Results and Discussion."

To further explore the relationships within these three clades a new scheme was devised. Forty-four genomes were selected from the clade, including all members corresponding to Newton's Clade 1, for inclusion in this step. As most ribosomal proteins are quite short, only 18 ribosomal genes were used and 15 single-copy housekeeping genes were added. This offered the advantage of adding a net of ~8,300 positions to the alignment, most of which are likely under less stringent selection than those of a ribosomal protein. An added advantage is that all 33 genes are shared with the Newton set. This creates a direct relationship facilitating comparison with that previous work. The top Blast hits for the 44 genomes were processed as described above for the ribosomal tree. The AICc test reported the same model for evolution as above. The tree was also generated using SPR and 100 bootstrap replicates. The resulting tree was rooted based on the ribosomal tree's placement of the clades. This corresponded to the root being placed where Newton's clades 1 and 2/4 diverge.

AVERAGE NUCLEOTIDE IDENTITY

JSpecies1.2.1 (Richter and Rosselló-Móra, 2009) was used as described previously (Fullmer et al., 2014) to analyze the genomes for average nucleotide identity (ANI) and tetramer frequency patterns.

SIDEROPHORE BIOCHEMICAL ASSAYS

To reduce contaminating iron, all glassware was washed and all solutions were prepared using water treated with a Nanopure Diamond filtration system (Barnstead, Lake Balboa, CA, USA). Siderophore production was confirmed using chrome azurol S (CAS) agar, modified for marine bacteria as previously described Whistler and Ruby (2003).

To test viability of ANG bacteria in iron-limiting conditions, several isolates were grown in the presence of the iron chelator ethylenediamine-N,N'-bis (2-hydroxyphenylacetic acid) (EDDHA) as described previously (McMillan et al., 2010). Cultures were grown for 24 h at 26°C in SWT then washed 3x in minimal sea salts solution (MSS, 50 mM $MgSO_4$, 10 mM $CaCl_2$, 350 mM NaCl, 10 mM KCl, 18.5 mM NH_4Cl, 333 μM K_2PO_4, $FeCl_3$ 10 μM, 100 mM PIPES, pH = 7.2) with no added iron or EDDHA. Cultures were inoculated to an OD_{600} of 0.05 in MSS with 10 μM $FeCl_3$. Glucose and casamino acids were added as carbon sources at 0.2 and 0.3% respectively and cultures were grown for 24 h at 26°C with shaking. To create iron-limiting conditions, EDDHA was added to the growth media at 10–30 μM. To test viability in iron-limiting conditions, cultures were grown for 24 h at 26°C, and the OD_{600} of each culture was measured and compared to control cultures without EDDHA. Siderophore

Table 1 | Genome assembly statistics for *Roseobacter* clade ANG isolates.

Isolate	Genome size (Mb)	# of genes	Missing genes* (% of total)	% GC	N50 (kb)	Contigs	Fold- coverage	Female ID
ANG-Vp	5.150	4,941	51 (1.0)	62.3	70	165	69.2	1
ANG-M1	5.375	5,097	63 (1.2)	62.0	211	180	132.3	3
ANG1	4.587	4,484	26 (0.6)	62.8	450	36	1,455†	1
ANG-DT	4.596	4,467	23 (0.5)	62.6	189	116	115.4	5
ANG-S	4.572	4,458	19 (0.4)	62.8	196	83	65.5	4
ANG-S3	4.597	4,468	18 (0.4)	62.7	300	84	129.0	2
ANG-M6	4.542	4,429	26 (0.6)	62.7	157	65	118.0	3
ANG-S5	4.660	4,534	33 (0.7)	62.5	233	54	123.5	2
ANG-M7	4.582	4,498	46 (1.0)	62.5	263	61	148.7	3
ANG-R	4.685	4,755	43 (0.9)	57.4	390	47	98.1	4
ANG-S4	4.538	4,619	9 (0.2)	57.2	978	20	71.9	2
ANG-S1	4.425	4,478	33 (0.7)	60.6	229	33	110.7	2

*As predicted by the RAST server (Aziz et al., 2008).
†*Leisingera* sp. ANG1 was previously sequenced with an Illumina mate-pair library and is therefore and a much higher fold coverage than other genomes (Collins and Nyholm, 2011).

production was measured from supernatants using the CAS liquid assay as described previously (Schwyn and Neilands, 1987). Further chemical characterization of siderophores was done using the Arnow (1937) and Csáky (1948) assays.

HOMOSERINE LACTONE DETECTION

Homoserine lactone (HSL) production was detected using the HSL-sensing bacterium *Agrobacterium tumefaciens* NTL4 (pZLR4; Cha et al., 1998). To determine acyl homoserine lactone (AHL) production, we used a well-diffusion assay as previously described Ravn et al. (2001). Briefly, a 3-mL culture of *A. tumefaciens* NTL4 was grown for 24 h in LB with gentamicin 30 μg/mL at 28°C. One milliliter of this culture was used to inoculate 50 mL of AB minimal media containing 0.5% glucose and 0.5% casamino acids (Chilton et al., 1974). After a 24-h incubation, 100 mL of AB minimal media containing 1.2% agar was autoclaved. Once the molten agar had cooled sufficiently, glucose and casamino acids were added to 0.5% each and 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) was added to a final concentration of 75 μg/mL. The molten agar was then combined with the 24-h culture of *A. tumefaciens*, distributed into petri dishes and allowed to solidify.

To induce HSL production by ANG isolates, cultures were grown overnight at 26°C in either SWT or MSS with 30 μM $FeCl_3$ and 0.5% of both glucose and casamino acids. To prevent the degradation of HSLs in alkaline conditions, the growth medium was buffered to pH 6.8 and never rose above 7.5 for any experiments. After a 24-h incubation, the cells were pelleted by centrifugation and the supernatant was filtered through a 0.22-μm filter. Wells were created in the *A. tumefaciens* agar plates using a sterile borer and 60 μL of cell-free supernatant was deposited into each well.

Accessory nidamental gland tissue was tested for the presence of AHLs by dissecting three separate ANGs from mature females as described above. Each ANG was homogenized in 300 μL of squid Ringer's solution and the homogenate was centrifuged at 1,000 × *g* for 10 min to pellet the ANG tissue. The supernatant containing bacterial cells was removed and centrifuged again at 10,000 × *g* for 10 min and 60 μL of the resulting clarified homogenate was deposited in a well of the AHL detection plates. All AHL detection plates were incubated at 28°C and photographed after 48 h.

RESULTS AND DISCUSSION

The genomes sequenced in this study were of a typical size for roseobacters, ranging from 4.4 to 5.4 Mb (**Table 1**). These large genomes are typical of the many cultured and sequenced organisms of the *Roseobacter* clade and reflect the diverse metabolisms reported in these bacteria (Newton et al., 2010). These data suggest that there has been little gene loss (or genome decay) as a result of close association with a host. However, several uncultivated roseobacters have streamlined genomes and may have a different lifestyle than most cultured members of this group (Luo et al., 2012, 2014). Many combinations of gene clusters for plasmid replication and partitioning were detected, particularly *repABC* genes. These data suggest that the ANG isolates have several extrachromosomal elements that may be resolved pending further sequencing efforts.

TAXONOMIC ANALYSIS

Of the ANG isolates identified, there were nine *Leisingera* (ANG1, ANG-DT, ANG-S, ANG-S3, ANG-S5, ANG-M6, ANG-M7, ANG-Vp, and ANG-M1), two *Ruegeria* (ANG-R and ANG-S4), and one *Tateyamaria* (ANG-S1) isolates. The 33 gene phylogenetic reconstruction placed these ANG isolates in five well-supported clades (**Figure 2**). The *Leisingera* isolates all grouped together in a single strongly supported clade, sister to four other described *Leisingera* taxa. This placement supports their recent designation as members of the *Leisingera* genus (Breider et al., 2014). The two

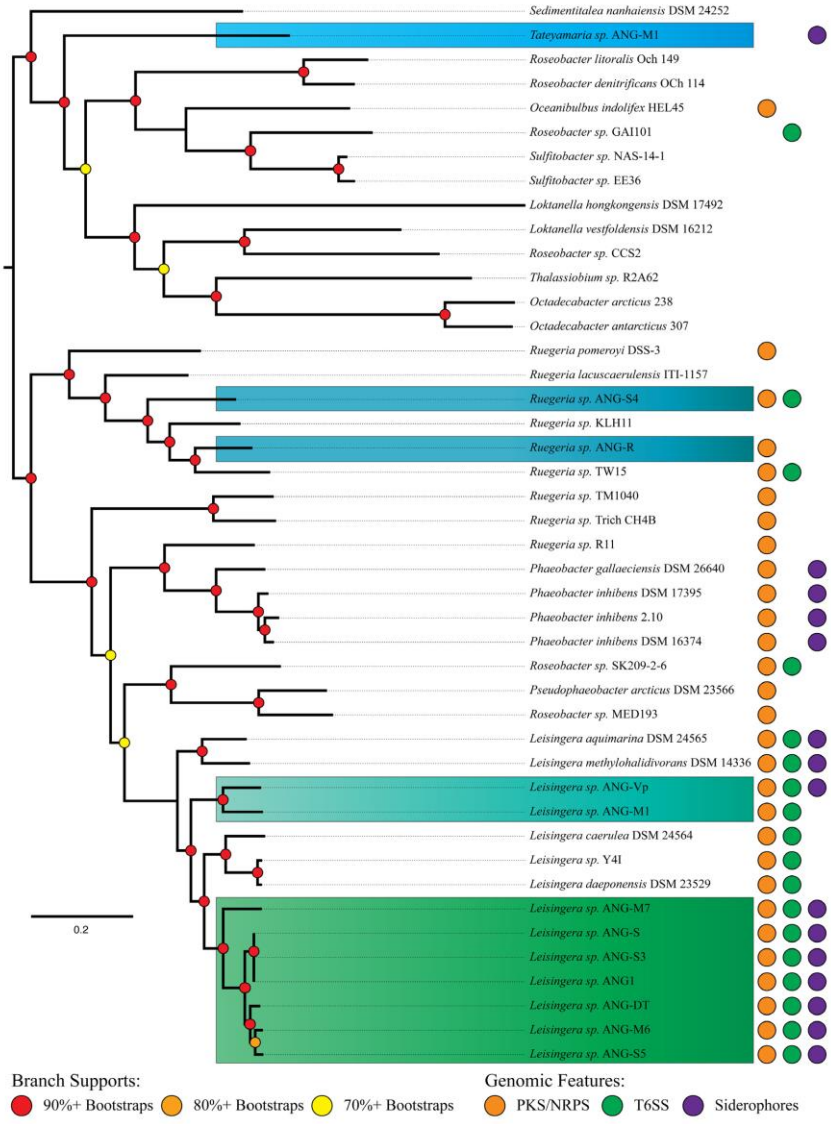


FIGURE 2 | MLSA analysis of *Roseobacter* clade isolates from the ANG with closely related organisms and distribution of significant gene clusters. Phylogenetic analysis of 33 single-copy housekeeping genes places most ANG isolates in the previously described "Clade 1" of the *Roseobacter* clade (Newton et al., 2010). A polyketide/non-ribosomal peptide synthase gene cluster is distributed throughout most of the *Roseobacter* clade, while Type VI secretion systems and siderophores are limited to fewer members. Isolates are color-coded to indicate genus. Blue, *Ruegeria*; Light blue, *Tateyamaria*; Teal and green, *Leisingera*.

Ruegeria ANG taxa did not place together, however, they are part of a clade composed of only *Ruegeria* taxa, affirming the putative genus designation. *Tateyamaria* placed on a basal branch long enough to suggest it is not closely associated with any of the taxa analyzed for this study.

The structure of the ribosomal tree (Supplmentary Figure 1) shares similarities with Newton et al.'s (2010) phylogeny. However, there are notable differences. First, the taxa of Newton's clade 3 are split into two separate clusters. Second, all but one member of Newton's clade 4 groups sister to clade 2. Finally, the two Rhodobacterales bacteria (HTCCs 2255 and 54623) fall among clades 2 and 4 rather than as part of the outgroups. The placement of clades 3 and 4 may be explained by the nature of gene concatenation. Concatenations can yield trees with high support values on topologies for which none of the constituents' gene phylogenies match (Salichos and Rokas, 2013; Colston et al., 2014). Gene choice can result in significantly different well-supported topologies. Thus, the averaged history of the ribosome may have been "outvoted" by the average history of the balance of Newton's seventy single-copy genes. The topology of the ribosomal tree was used to assign the root in the 33 gene tree (**Figure 2**) on the assumption that the ribosomal phylogeny was accurate in clade 4's placement. The clade 4 taxa could be used as outgroups to clades 1 and 2 instead with no significant change to the further analyses of the ANG isolates.

The structure of the 33 gene tree (**Figure 2**) compares well with Newton's phylogeny. Taxa previously identified as *Phaeobacter, Ruegeria, and Leisingera* formed polyphyletic clades. This occurrence was not unanticipated as the Newton et al. (2010) study showed a 70-gene tree with the same structure, albeit with fewer taxa. The genes analyzed in this study represent a subset of those analyzed in Newton et al. (2010) and therefore were expected to recapitulate this result. Our tree also aligns well with the recent reclassification by Breider et al. (2014). *Sedimentitalea nanhaiensis*, formerly *Leisingera nanhaiensis*, placed at the base of Newton's clade 2, which is separated from the balance of the *Leisingera* genus. *Pseudophaeobacter arcticus*, formerly *Phaeobacter arcticus*, fell in a clade sister to the *Leisingera*, also isolated from the newly redefined *Phaeobacter* genus. Thus, its reclassification resolves a polyphyly observed in our tree. Likewise, *L. caerulea* and *L. daeponensis*, also reclassified from the *Phaeobacter* genus, resolve a separate polyphyly. As these two taxa are sister to established *Leisingera*, we find reassigning them to this genus in line with our results. The only remaining question of polyphyly in our 33 gene phylogeny is *Ruegeria* sp. R11, which groups with the *Phaeobacter/Pseudophaeobacter/Leisingera* clade. This isolate has been proposed as *Nautella* based on 16S rDNA similarity to the *Nautella* type strain and may not be a member of the *Ruegeria* genus (Fernandes et al., 2011).

The phylogenetic analyses identified apparent relationships at approximately the genus level. In order to attempt to refine these results and provide species-level putative designations, ANI was employed using the accepted ANI cutoff of 95% (**Figure 3**, Konstantinidis et al., 2006; Richter and Rosselló-Móra, 2009). The ANG isolates fall into seven putative species groups. Six *Leisingera* isolates (ANG1, ANG-DT, ANG-S, ANG-S3, ANG-S5, and ANG-M6) formed one group, and three other *Leisingera* isolates (ANG-M7, ANG-M1, ANG-Vp) and the two *Ruegeria* isolates each formed its own singleton group. The *Leisingera* isolates are of particular interest as previous research has shown that the most common symbionts within the ANG belong to the genus *Leisingera*, though they were previously classified within the genus *Phaeobacter* (Collins and Nyholm, 2011; Collins et al., 2012). One putative species of *Leisingera* was consistently isolated from the five individual ANGs used in this study. This cluster of isolates likely represents the dominant culturable symbiont present in the ANG and includes the previously sequenced isolate, *Leisingera* sp. ANG1. Notably, the ANI values of the ANG isolates all fell short of even 90% identity with any of the previously described species. These data suggest that each of these putative ANG species is, indeed, a novel taxon. Providing comprehensive polyphasic species descriptions is beyond the scope of this work, so we propose these taxa as sp. of their various assigned genera.

RECLASSIFICATION OF *Phaeobacter gallaeciensis* ANG1

Consistent with previous research, our results suggest the isolate we had previously identified as *P. gallaeciensis* is phylogenetically distinct from the type species, *P. gallaeciensis* DSM 26640 (Thole et al., 2012; Breider et al., 2014). We therefore reclassify the isolate *P. gallaeciensis* ANG1 as *Leisingera* sp. ANG1 pending further phenotypic analyses.

GENOME CHARACTERISTICS AND GENERAL METABOLISM

Of the 12 ANG symbionts examined in this study, all have genes encoding a complete Entner–Doudoroff pathway for metabolizing glucose. Furthermore, all of them lack the gene for phosphofructokinase, a key enzyme from the Embden–Meyerhof–Parnas pathway. This is typical of many previously sequenced and complete genomes from the *Roseobacter* lineage (Moran et al., 2004; Newton et al., 2010; Wagner-Döbler et al., 2010). Two organisms (*Tateyamaria* sp. ANG-S1 and *Ruegeria* sp. ANG-S4) contain all genes for a complete pentose-phosphate pathway. The others contain most genes for the pathway, with the exception of a gene encoding 6-phosphogluconate dehydrogenase. As an alternative metabolic pathway, 6-phosphogluconate produced by the first two enzymes of the pentose phosphate pathway could feed into the Entner–Doudoroff pathway for further carbohydrate metabolism (Fuchs, 1999; Berger et al., 2014).

While the *Roseobacter* clade was first described as a group of obligate aerobic organisms, recently it has been shown that some members contain enzymes needed for anaerobic respiration of nitrate (Dogs et al., 2013). All of the isolates from the ANG contain the gene for nitrate reductase that could be used for anaerobic respiration of nitrogen. Most isolates, with the exception of *Tateyamaria* sp. ANG-S1, also contain genes for other denitrifying enzymes to further reduce nitrogenous oxyanions. These data suggest that the ANG isolates may be able to survive and thrive in anaerobic environments by respiring nitrogenous oxyanions.

Although genes associated with phototrophy were detected in *Tateyamaria* ANG-S1, including bacteriochlorophyll a, these genes were not detected in the other ANG isolates. These data are consistent with previous observations of Clade-1 roseobacters which

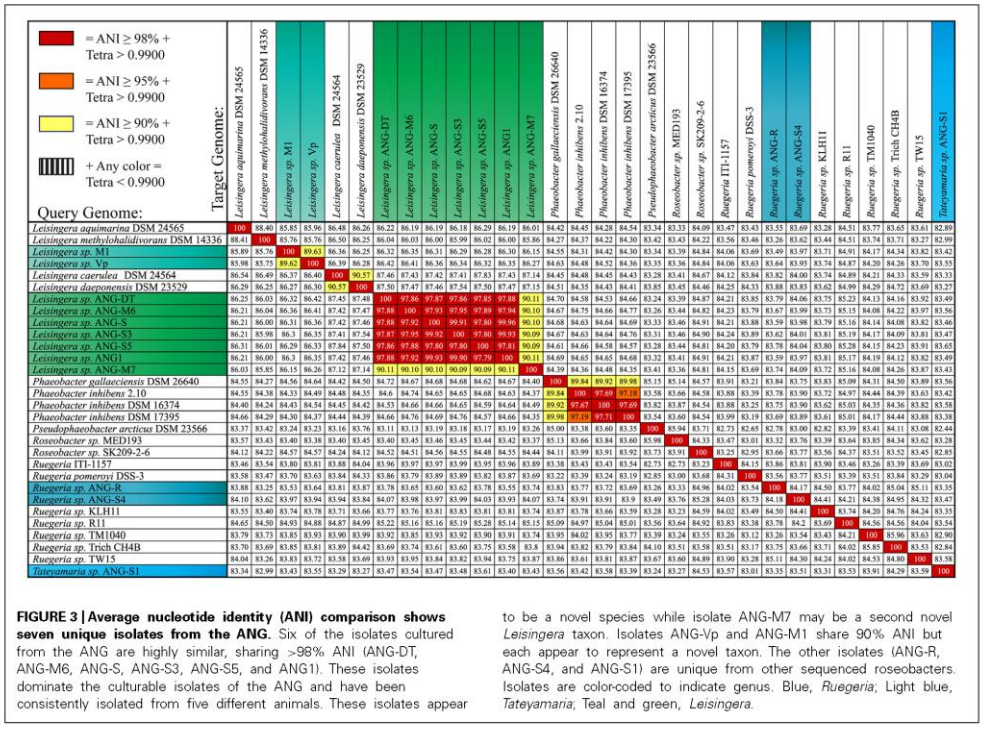


FIGURE 3 | Average nucleotide identity (ANI) comparison shows seven unique isolates from the ANG. Six of the isolates cultured from the ANG are highly similar, sharing >98% ANI (ANG-DT, ANG-M6, ANG-S, ANG-S3, ANG-S5, and ANG1). These isolates dominate the culturable isolates of the ANG and have been consistently isolated from five different animals. These isolates appear to be a novel species while isolate ANG-M7 may be a second novel *Leisingera* taxon. Isolates ANG-Vp and ANG-M1 share 90% ANI but each appear to represent a novel taxon. The other isolates (ANG-R, ANG-S4, and ANG-S1) are unique from other sequenced roseobacters. Isolates are color-coded to indicate genus. Blue, *Ruegeria*; Light blue, *Tateyamaria*; Teal and green, *Leisingera*.

were not found to be phototrophic (Newton et al., 2010; Luo and Moran, 2014).

PROTEIN SECRETION SYSTEMS

While a Type IV secretion system is present in many roseobacters, we detected *virB* in only two of the genomes examined here (ANG-M1 and ANG-R). Previous literature has suggested these systems are used for communication between bacteria and eukaryotic cells (Luo and Moran, 2014). However, given that a large proportion of isolates from the ANG appear to lack this system, the T4SS may not be a critical means of communication between the consortium and its host.

An interesting feature of the *Leisingera* genus is that all sequenced genomes contain genes for a Type VI secretion system (T6SS, **Figure 2**). In *L. daeponensis* and *L. caerulea* it has been shown that this T6SS exists on a plasmid (Beyersmann et al., 2013; Dogs et al., 2013). In *Leisingera* sp. ANG1 the T6SS is located on a large contig (>500 kb) containing *repAB* plasmid partitioning genes, suggesting that the T6SS in this species is also located on a plasmid. Similar conclusions were reached with the genomes of *L. caerulea, L. daeponensis, L. methylohalidivorans,* and *L. aquimarina.* Each of these organisms has genes for a T6SS on plasmids that vary in size (from 109 kB in *L. caerulea* to 526 kB in *Leisingera* sp. ANG1); however, all have a DnaA 1-like replicase (Beyersmann et al., 2013; Buddruhs et al., 2013; Dogs et al., 2013; Riedel et al., 2013). While other roseobacters contain a T6SS, the conservation of the T6SS on similar plasmids could be characteristic of this genus.

Several functions of the T6SS have been proposed, including antimicrobial roles, as evidenced by direct cell-contact mediated killing (Murdoch et al., 2011; Russell et al., 2011). The T6SS has also been shown to be involved with host–microbe interactions, particularly in the *Rhizobiales. A. tumefaciens* shows attenuated ability to create crown gall tumors when the T6SS is deleted (Wu et al., 2008). Similarly, the nitrogen-fixing plant symbiont *Rhizobium leguminosarum* lacking a T6SS will successfully colonize its host, however, it will fail to fix nitrogen (Bladergroen et al., 2003). The T6SS has also been implied in many other general associations between microorganisms, including predator evasion (Pukatzki et al., 2006) and self/non-self recognition (Gibbs et al., 2008).

It is interesting that all of the isolates, with one exception (*Ruegeria* sp. ANG-R), have genes for a T6SS, including isolates outside of the *Leisingera* genus. This suggests that the T6SS in these bacteria may be important for communication with the host and/or with other bacteria. In the ANG of *E. scolopes,* bacteria are housed in high densities within the epithelium-lined tubules

of the organ (Collins et al., 2012). Such high densities of bacterial cells foster close contact with other bacteria and many host cells, including the ANG epithelium and hemocytes, the principle cellular innate immunity component of the host. Given that the T6SS functions by direct cell-to-cell contact, it would be an ideal mechanism for the delivery of effectors directly to other symbionts and/or host tissues. The T6SS may play a role in mediating how these organisms are selected from the environment and explain how some species are able to dominate the bacterial populations within a given tubule (Collins et al., 2012).

SECONDARY METABOLITES

Members of the *Roseobacter* clade have been shown to produce several unique secondary metabolites. Some of the most notable ones include antibacterials such as TDA, produced by organisms such as *P. inhibens* and *Ruegeria* sp. TM1040, and the blue pigment indigoidine, produced by organisms such as *Leisingera* sp. Y4I and *L. daeponensis* (Geng et al., 2008; Cude et al., 2012). None of the biosynthetic genes for either of these compounds were found in any of the genomes sequenced. Furthermore, no classical antibiotic synthesis pathways (e.g., tetracycline, carbapenems, etc.) were found.

However, analysis with the Antibiotic and Secondary Metabolite Analysis Shell (AntiSMASH, Blin et al., 2013) revealed several gene clusters encoding potential secondary metabolism (**Table 2**). These included gene clusters for siderophore synthesis, autoinducer synthases (*luxI* homologs), polyketide/non-ribosomal peptide synthases (PKS/NRPS) and production of volatile compounds such as terpenes. The BActeriocin GEnome mining tooL (BAGEL, van Heel et al., 2013), was used to screen genomes for possible bacteriocin producing gene clusters, which were found in the *Ruegeria* isolates (ANG-R and ANG-S4) as well as *Tateyamaria* sp. ANG-S1 (**Table 2**). Bacteriocins are a broad group of proteins that can be used to kill other bacteria but have also been shown to act as inducers of invertebrate metamorphosis and thus may serve a number of functions (Cotter et al., 2013; Shikuma et al., 2014).

All isolates have a conserved non-ribosomal peptide/polyketide synthase gene cluster characterized previously (**Table 2**, Martens et al., 2006). This gene cluster is conserved in the *Roseobacter* lineage, being found in 28 of 57 previously sequenced genomes, and is comprised of four genes: a non-ribosomal polypeptide synthase, a polyketide synthase, a glycosyltransferase and a phosphopantetheinyl transferase. However, the product of this gene cluster has not yet been characterized. Given that this gene cluster is well-conserved throughout the *Roseobacter* lineage, its product and function should be elucidated through future experiments.

QUORUM SENSING

Homoserine lactones produced by LuxI homologs have been widely studied as quorum sensing molecules in bacteria, including the *luxIR* system of *V. fischeri*, the light organ symbiont of *E. scolopes* (Antunes et al., 2007; Miyashiro and Ruby, 2012). AntiSMASH detected 2 separate pairs of *luxIR* homologs in the ANG isolates that were most similar to the *ssaIR* and *ssbIR* previously described in *Ruegeria* sp. KLH11 (Zan et al., 2012). However, only the *Ruegeria* isolates, ANG-S4 and ANG-R, have both pairs of *luxIR* homologs. Most of the ANG bacteria only have homologs of *ssbIR*. In *Ruegeria* sp. KLH11, these two systems work together to control biofilm formation and motility (Zan et al., 2012). The genes *ssaI* and *ssaR*, are shown to regulate the change between adherent and planktonic lifestyles. Increased levels of HSLs promote flagellar growth and motility, while lower levels foster biofilm development. The actions of these genes can be indirectly repressed by *ssbIR*. The fact that so many ANG isolates have only the *ssbIR* homologs suggest that there may be a unique function for these quorum sensing genes independent of the *ssaIR* quorum sensing system. In addition to *ssbIR*, the ubiquitous *luxIR* homologs in the *Roseobacter* genomes from the ANG are also similar to the *raiIR* genes described in *Rhizobium etli* (Rosemeyer et al., 1998). Both SsbIR and RaiIR are known to produce 3-hydroxyl-HSL compounds, but *raiIR* has been shown to control growth and nitrogen fixation, not motility. This raises the possibility that the *luxIR* genes in ANG roseobacters may regulate growth of bacteria within the ANG.

Table 2 | Secondary metabolite gene clusters detected with AntiSMASH and BAGEL.

	PKS/NRPS	LuxRI	Bacteriocin	Siderophore	Terpene	Ectoine
Leisingera sp. ANG-Vp	1	1	0	1	0	1
Leisingera sp. ANG-M1	1	1	0	0	0	1
Leisingera sp. ANG1	1	1	0	1	0	0
Leisingera sp. ANG-DT	1	1	0	1	0	0
Leisingera sp. ANG-S	1	1	0	1	0	0
Leisingera sp. ANG-S3	1	1	0	1	0	0
Leisingera sp. ANG-M6	1	1	0	1	0	0
Leisingera sp. ANG-S5	1	1	0	1	0	0
Leisingera sp. ANG-M7	1	1	0	1	0	0
Ruegeria sp. ANG-R	1	2	3	0	0	1
Ruegeria sp. ANG-S4	2	2	3	0	0	0
Tateyamaria sp. ANG-S1	0	1	2	1	1	0

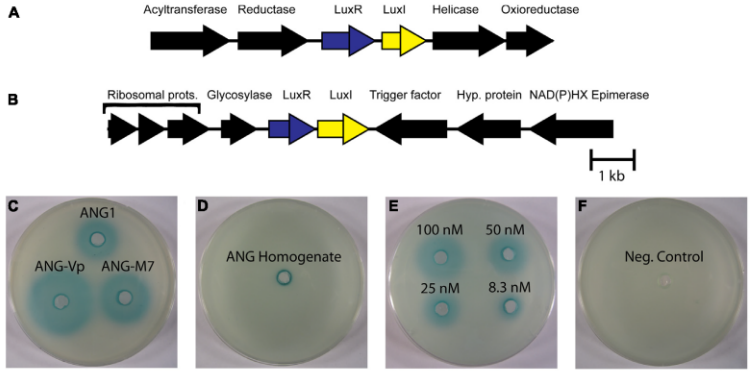


FIGURE 4 | LuxIR homologs in *Roseobacter* clade organisms from the ANG and associated homoserine lactone production. (A) All ANG isolates have a pair of *luxIR* homologs flanked by potential anabolic genes (crontonyl CoA reductase and acetyltransferase) and a helicase and oxioreductase. **(B)** Another pair of *luxIR* homologs is present only in the *Ruegeria* species (isolates ANG-S4 and ANG-R) flanked by ribosomal proteins, a cell division trigger factor and a glycosylase. **(C)** Representative picture of β-galactosidase activity induced by homoserine lactone in supernatants from cultures of *Leisingera* sp. ANG1, *Leisingera* sp. ANG-M7, and *Leisingera* sp. ANG-Vp. **(D)** Homoserine lactones were also detected in ANG homogenates (representative image, $n = 3$ separate ANGs). **(E)** Semi-quantitative dilution of N-3-oxo-hexanoyl-homoserine lactone. **(F)** 60 μL growth medium (Negative control).

To determine if HSLs are present in the ANG and are produced by the bacterial symbionts, we tested for the presence of AHLs using a semi-quantitative biosensor assay. All isolates that could grow to high density in liquid medium produced detectable HSLs (**Figure 4**). Species like *Tateyamaria* sp. S1 did not grow to a very high density and failed to produce enough HSL to be detected by the assay (not shown). The homogenates of three ANGs were also tested and resulted in small zones of β-galactosidase activity around the assay wells, suggesting that HSLs are produced in the ANG and could contribute to the symbiosis by influencing gene expression of the bacterial consortium. As a negative control, host gill tissue was also homogenized in a similar manner to ensure that compounds from squid tissue were not inducing expression of β-galactosidase in the *A. tumefaciens* biosensor. No enzymatic activity was observed in this control (not shown), confirming the specificity of the assay.

While HSLs were detected in both pure culture and in ANG homogenate, gene regulation by HSL quorum sensing may be different than what has been described for their nearest homologs in *Ruegeria* sp. KLH11. Most ANG isolates, including the dominant *Leisingera* species, lack the *ssaIR* homologs directly responsible for the increase of motility described in *Ruegeria* sp. KLH11. This suggests there is a yet undescribed role for the *ssbIR* homologs in the *Roseobacter* clade isolates from the ANG.

Future research should investigate the chemical nature of the HSL produced by the autoinducer synthases in individual ANG isolates. The nearest characterized homologs, both RaiI in *R. etli* and SsbI in *Ruegeria* sp. KLH11 produce 3-hydroxyl-HSL compounds (Rosemeyer et al., 1998; Zan et al., 2012). Future studies should confirm that the members of the ANG consortium also produce this type of HSL. Genetics have not yet been developed in any of the cultured ANG isolates, however, creating a non-functioning mutant of the autoinducer synthase could reveal phenotypes controlled by quorum sensing. Comparing transcriptomes between HSL^- and wild type strains may also reveal genes that are controlled by quorum sensing. Moreover, previous research has shown that these symbionts are likely environmentally transmitted (Kaufman et al., 1998). Thus, the symbionts encounter three environments of varying cell density, from ambient seawater with a low density of symbionts, to the tubules of the ANG where the cells are highly concentrated, to the egg jelly coat with a lower density. Given the profound differences in cell density between free-living symbionts in seawater and the tubules of an ANG, quorum sensing may be an ideal mechanism for gene regulation between the different environments experienced by ANG bacteria (host/ANG, egg, free-living). Further studies should also examine how gene expression changes from the high-cell density environment of the ANG to the egg jelly coat, where cell densities will be lower, but where any anti-fouling compounds may be produced.

SIDEROPHORES

Another group of secondary metabolite biosynthesis genes that was detected in the genomes of ANG isolates were siderophores. Siderophores are small molecules with high affinities for iron and can be used by bacteria for iron scavenging. Iron is needed for many cellular functions, including respiration, detoxification of reactive oxygen species (e.g., catalases, super-oxidase dismutase), and metabolism (e.g., aconitase of the TCA cycle). Very few organisms are known to survive without iron (Andrews et al., 2003). One way that bacteria can acquire iron in environments where it is a limiting resource is by producing siderophores to sequester iron from other sources.

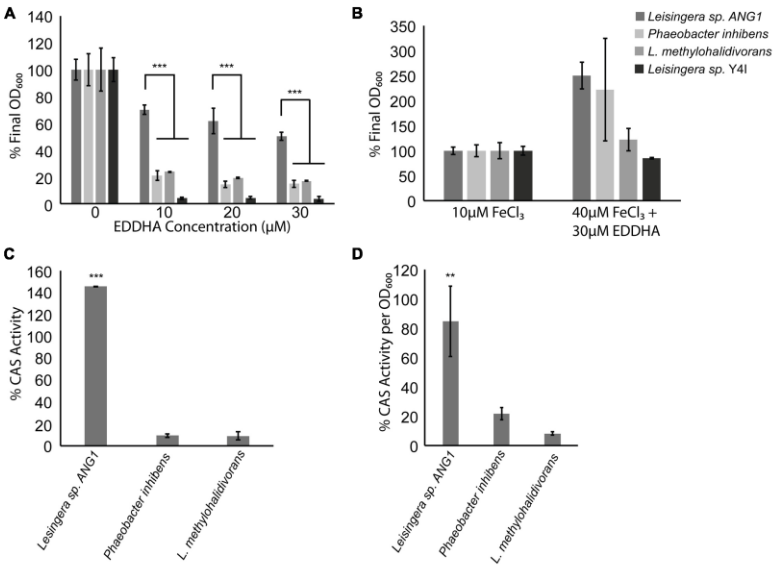


FIGURE 5 | *Roseobacter* clade symbionts from the ANG have a growth advantage in iron-limiting conditions, possibly due to siderophore production. (A) While other *Roseobacter* clade organisms were inhibited by the presence of an iron-chelator, *Leisingera* sp. ANG1 grew to more than 50% of its optical density even if EDDHA was at three times the concentration of available iron. **(B)** EDDHA is not toxic to *Roseobacter* clade organisms, as adding enough iron to overwhelm the chelator restored the growth defect to all *Roseobacter* clade organisms. **(C)** *Leisingera* sp. ANG1 produced more siderophore than non-ANG isolates in the presence of EDDHA. **(D)** Even if no iron chelator were present, siderophores were more abundant in the supernatants from cultures of *Leisingera* sp. ANG1. ** $p < 0.01$, *** $p < 0.001$.

Siderophore synthesis genes in the *Roseobacter* clade are rare. Of previously sequenced *Roseobacter* genomes, only six genomes from four species (*L. aquimarina*, *L. methylohalidivorans*, *P. inhibens*, and *P. gallaeciensis*) are predicted to have siderophore synthesis genes (**Figure 2**). However, all roseobacters isolated from the ANG of *E. scolopes*, with the exception of *Ruegeria* sp. ANG-R and ANG-S4, have either siderophore biosynthesis genes or showed siderophore activity in biochemical assays (**Table 2** and **Figure 5**). For example, *Leisingera* sp. ANG-M1 had no predicted siderophore synthesis genes, but siderophore activity was detected when grown on CAS agar and in CAS liquid assays, suggesting that these biosynthetic genes may not be annotated, perhaps due to the fragmented state of the assembled genome for this isolate. Conversely, *Tateyamaria* sp. ANG-S1 has siderophore biosynthetic genes, but failed to show siderophore activity (not shown). Taken together, these data suggest induction of siderophore synthesis genes may be controlled very differently in *Tateyamaria* sp. ANG-S1 and may be induced only under specific conditions.

We compared growth and siderophore production in iron-limiting conditions of *Leisingera* sp. ANG1, a representative of the dominant ANG symbionts, to three other species from the *Roseobacter* lineage. Siderophore-producing strains *P. inhibens* DSMZ 17395 and *L. methylohalidivorans* DSM 14336 were tested along with the non-siderophore producing strain *Leisingera* sp. Y4I. When grown in the presence of the iron chelator EDDHA, most roseobacters had a growth defect, growing to only 20% of the control density (**Figure 5A**). However, *Leisingera* sp. ANG1 had a much smaller growth defect ($p < 0.001$), growing to greater than 50% of the control OD when concentrations of EDDHA were three times the concentration of available iron in the media (**Figure 5A**). To show this was not due to a toxic effect of EDDHA, $FeCl_3$ was added to higher concentrations (40 μM) to overwhelm the iron chelator, which restored the growth of all organisms (**Figure 5B**).

The survival of *Leisingera* sp. ANG1 under iron-limiting conditions could be due to the higher levels of siderophores produced by these organisms. Supernatants from cultures of strains that failed to grow (*P. inhibens* and *L. methylohalidivorans*) showed very little CAS activity while supernatants from cultures of ANG1 had very high levels of CAS activity, indicative of a high concentration of siderophores ($p < 0.001$, **Figure 5C**). To determine if this increase was a consequence of the increased growth of *Leisingera* ANG1, CAS activity was measured in supernatants from cultures without any iron chelator added. This allowed the bacteria to grow and deplete the iron available in the media, leading to induction of

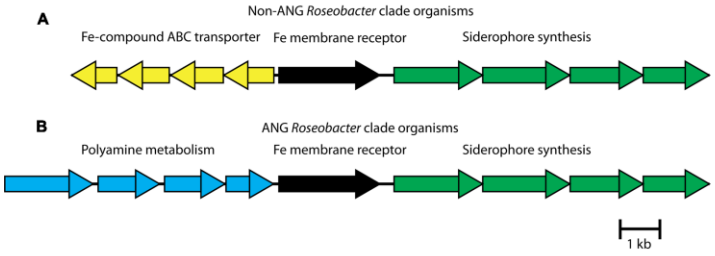


FIGURE 6 | *Roseobacter* isolates from the ANG have a unique genome rearrangement upstream of siderophore biosynthesis group. (A) In previously sequenced *Leisingera* and *Phaeobacter* species, an ABC transporter, predicted to transport iron compounds like iron dictate, lies upstream of a membrane receptor protein and the siderophore synthesis gene cluster. **(B)** In ANG bacteria, polyamine metabolism genes replace the ABC transporter genes. In addition to modifying regulatory elements, the four polyamine genes could synthesize a polyamine backbone of the siderophores, such as putrescine.

siderophore synthesis. Supernatants from cultures of *Leisingera* sp. ANG1 had more CAS activity than either *P. inhibens* DSM17395 or *L. methylohalidivorans* DSM14336 per unit OD_{600} ($p < 0.01$, **Figure 5D**). These data suggest that the abundance of siderophores produced by *Leisingera* sp. ANG1 is not just due to an increase in cell number, but instead to increased siderophore production at the cellular level.

Examining the siderophore biosynthesis genes in roseobacters isolated from the ANG, revealed a unique genome rearrangement (**Figure 6**). In all other siderophore-producing roseobacters, siderophore synthesis genes are located downstream of an iron membrane receptor and an iron-compound ABC transporter. In roseobacters isolated from the ANG, four genes related to polyamine metabolism are inserted upstream of the iron membrane receptor (**Figure 6**). The polyamine genes upstream of the siderophore synthesis cluster are sufficient to synthesize putrescine, a backbone of certain catechol siderophores such as photobactin from *Photorhabdus luminescens* (Ciche et al., 2003). Testing the supernatant of *Leisingera* sp. ANG1 with the Arnow assay showed that catechol siderophores were being produced. This genome rearrangement may be responsible for the higher production of siderophores in *Leisingera* sp. ANG1, perhaps by altering the regulatory elements upstream of the siderophore biosynthesis genes or perhaps by coupling the production of putrescine and the catecholate siderophore. Future research may determine if putrescine is a structural component of the catechol siderophores produced by the ANG symbionts such as the dominant *Leisingera* symbionts.

Producing siderophores can be beneficial to bacteria that colonize animal tissues. Iron-chelating proteins produced by hosts can effectively deplete freely available iron to the associated microbiota (Ong et al., 2006). Furthermore, a host infected with a pathogen will sometimes increase production of iron-chelating proteins as a way to starve infectious bacteria of a critical resource (Jurado, 1997). One of the most-widely studied models is the siderophore enterobactin which is produced by several species of enteric bacteria, including *Salmonella* and *Escherichia* species (Raymond et al., 2003). This iron-chelating molecule acquires iron from serum proteins carrying iron, such as transferrin, and the siderophore-iron complex is taken up by the infecting bacteria to keep them supplied with iron. To combat this, the innate immune system produces proteins to bind siderophores in order to prevent the iron-scavenging molecules from fulfilling their purpose (Goetz et al., 2002; Abergel et al., 2008).

In invertebrates, iron sequestration can be performed by two ubiquitous proteins, ferritin, and transferrin. Ferritin is present in the hemolymph of invertebrates where it can function as an iron transporter or iron scavenger (Ong et al., 2005) and transferrin is up-regulated in insect epithelia during bacterial infection (Buchon et al., 2009; Wang et al., 2009). Both of these proteins have been found in transcriptomic and proteomic data from both hemocytes and light organ tissues of *E. scolopes* (Schleicher and Nyholm, 2011, Collins, unpublished data). These iron chelators, if present in the ANG, could provide a selective pressure that other roseobacters would have to overcome. In such a case, siderophore-producing organisms such as *Leisingera* sp. ANG1 may have an advantage over other bacteria and this may contribute to its dominance in the consortium. Colonization of cephalopod ANGs is likely via environmental transmission (Kaufman et al., 1998) and overcoming iron-limitation may be one part of what is likely a complex process for establishment and development of the association.

The function of the ANG and its bacterial consortium remains unknown even though it was hypothesized that the bacteria deposited in the jelly coats of squid eggs may play a role in protecting the egg masses from fouling, possibly through the production of antimicrobial compound(s) (Biggs and Epel, 1991). Previous research in the eggs of the shrimp (*Palaemon macrodactylus*) have shown that, once the eggs are brooded, *Alteromonas* sp. bacteria colonize the surface of the egg and produce the antimicrobial compound 2, 3-indolinedione that protects the eggs from fungal infection (Gil-Turnes et al., 1989). However, shrimp eggs acquire these epibionts from seawater which is an important distinction from squid eggs, where the bacterial symbionts from the ANG are actively deposited into jelly coat layers. Future research will attempt to understand the role of

these bacteria within the eggs of developing embryos and try to discern what contribution they may make to deter fouling organisms.

This study sets the foundation for future research on the ANG symbionts by characterizing the genomes of several isolates from the *Roseobacter* lineage. We have identified many features of these genomes that may be important in the ANG association including Type VI secretion systems, siderophore production and putative quorum sensing systems using HSLs. The ANG and associated roseobacters are found worldwide in many different cephalopod species. This trend suggests that the consortium may play a similar and conserved role in squid and cuttlefish. Future research will hopefully elucidate the contribution of these bacteria to the development and survival of cephalopods and their embryos. Genome analyses of the *Roseobacter* clade bacteria that dominate the ANG, along with future genomic and transcriptomic studies of other ANG symbionts and the entire consortium will provide a number of exciting avenues of research to help elucidate the nature of this widely distributed association.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Alison Buchan, Dr. Jeffra Schaefer, Dr. Stephen Farrand, and Dr. Mary Ann Moran for providing bacterial strains as well as the UConn Bioinformatics Facility for providing computing resources. This research was funded by NSF IOS-0958006 and the University of Connecticut Research Foundation to SVN.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2015.00123/abstract

REFERENCES

Abergel, R. J., Clifton, M. C., Pizarro, J. C., Warner, J. A., Shuh, D. K., Strong, R. K., et al. (2008). The siderocalin/enterobactin interaction: a link between mammalian immunity and bacterial iron transport. *J. Am. Chem. Soc.* 130, 11524–11534. doi: 10.1021/ja803524w

Andrews, S. C., Robinson, A. K., and Rodríguez-Quiñones, F. (2003). Bacterial iron homeostasis. *FEMS Microbiol. Rev.* 27, 215–237. doi: 10.1016/S0168-6445(03)00055-X

Antunes, L. C. M., Schaefer, A. L., Ferreira, R. B. R., Qin, N., Stevens, A. M., Ruby, E. G., et al. (2007). Transcriptome analysis of the *Vibrio fischeri* LuxR-LuxI regulon. *J. Bacteriol.* 189, 8387–8391. doi: 10.1128/JB.00736-07

Arnow, L. (1937). Colorimetric determination of the components of 3, 4-dihydroxyphenylalaninetyrosine mixtures. *J. Biol. Chem.* 118, 531–537.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75

Barbieri, E., Paster, B., Hughes, D., Zurek, L., Moser, D., Teske, A., et al. (2001). Phylogenetic characterization of epibiotic bacteria in the accessory nidamental gland and egg capsules of the squid *Loligo pealei* (Cephalopoda:Loliginidae). *Environ. Microbiol.* 3, 151–167. doi: 10.1046/j.1462-2920.2001.00172.x

Berger, A., Dohnt, K., Tielen, P., Jahn, D., Becker, J., and Wittmann, C. (2014). Robustness and plasticity of metabolic pathway flux among uropathogenic isolates of *Pseudomonas aeruginosa*. *PLoS ONE* 9:e88368. doi: 10.1371/journal.pone.0088368

Beyersmann, P. G., Chertkov, O., Petersen, J., Fiebig, A., Chen, A., Pati, A., et al. (2013). Genome sequence of *Phaeobacter caeruleus* type strain (DSM 24564(T)), a surface-associated member of the marine *Roseobacter* clade. *Stand. Genomic Sci.* 8, 403–419. doi: 10.4056/sigs.3927623

Biggs, J., and Epel, D. (1991). Egg capsule sheath of *Loligo opalescens* Berry: structure and association with bacteria. *J. Exp. Zool.* 259, 263–267. doi: 10.1002/jez.1402590217

Bladergroen, M. R., Badelt, K., and Spaink, H. P. (2003). Infection-blocking genes of a symbiotic *Rhizobium leguminosarum* strain that are involved in temperature-dependent protein secretion. *Mol. Plant Microbe Interact.* 16, 53–64. doi: 10.1094/MPMI.2003.16.1.53

Blin, K., Medema, M. H., Kazempour, D., Fischbach, M. A., Breitling, R., Takano, E., et al. (2013). antiSMASH 2.0 – a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res.* 41, W204–W212. doi: 10.1093/nar/gkt449

Bloodgood, R. (1977). The squid accessory nidamental gland: ultrastructure and association with bacteria. *Tissue Cell* 9, 197–208. doi: 10.1016/0040-8166(77)90016-7

Breider, S., Scheuner, C., Schumann, P., Fiebig, A., Petersen, J., Pradella, S., et al. (2014). Genome-scale data suggest reclassifications in the *Leisingera-Phaeobacter* cluster including proposals for *Sedimentitalea* gen. nov. and *Pseudophaeobacter* gen. nov. *Front. Microbiol.* 5:416. doi: 10.3389/fmicb.2014.00416

Brinkhoff, T., Bach, G., Heidorn, T., Liang, L., Schlingloff, A., and Simon, M. (2004). Antibiotic production by a *Roseobacter* clade-affiliated species from the German Wadden Sea and its antagonistic effects on indigenous isolates. *Appl. Environ. Microbiol.* 70, 2560–2565. doi: 10.1128/AEM.70.4.2560

Buchon, N., Broderick, N. A., Poidevin, M., Pradervand, S., and Lemaitre, B. (2009). *Drosophila* intestinal response to bacterial infection: activation of host defense and stem cell proliferation. *Cell Host Microbe* 5, 200–211. doi: 10.1016/j.chom.2009.01.003

Buddruhs, N., Chertkov, O., Petersen, J., Fiebig, A., Chen, A., Pati, A., et al. (2013). Complete genome sequence of the marine methyl-halide oxidizing *Leisingera methylohalidivorans* type strain (DSM 14336(T)), a representative of the *Roseobacter* clade. *Stand. Genomic Sci.* 9, 128–141. doi: 10.4056/sigs.4297965

Case, R. J., Longford, S. R., Campbell, A. H., Low, A., Tujula, N., Steinberg, P. D., et al. (2011). Temperature induced bacterial virulence and bleaching disease in a chemically defended marine macroalga. *Environ. Microbiol.* 13, 529–537. doi: 10.1111/j.1462-2920.2010.02356.x

Caspi, R., Altman, T., Dale, J. M., Dreher, K., Fulcher, C. A., Gilham, F., et al. (2010). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 38, D473–D479. doi: 10.1093/nar/gkp875

Cha, C., Gao, P., Chen, Y. C., Shaw, P. D., and Farrand, S. K. (1998). Production of acyl-homoserine lactone quorum-sensing signals by gram-negative plant-associated bacteria. *Mol. Plant Microbe Interact.* 11, 1119–1129. doi: 10.1094/MPMI.1998.11.11.1119

Chilton, M. D., Currier, T. C., Farrand, S. K., Bendich, A. J., Gordon, M. P., and Nester, E. W. (1974). *Agrobacterium tumefaciens* DNA and PS8 bacteriophage DNA not detected in crown gall tumors. *Proc. Natl. Acad. Sci. U.S.A.* 71, 3672–3676. doi: 10.1073/pnas.71.9.3672

Ciche, T. A., Blackburn, M., Carney, J. R., and Ensign, J. C. (2003). Photobactin: a catechol siderophore produced by *Photorhabdus luminescens*, an entomopathogen mutually associated with *Heterorhabditis bacteriophora* NC1 nematodes. *Appl. Environ. Microbiol.* 69, 4706–47013. doi: 10.1128/AEM.69.8.4706

Collins, A., LaBarre, B., Won, B., Shah, M., Heng, S., Choudhury, M., et al. (2012). Diversity and partitioning of bacterial populations within the accessory nidamental gland of the squid *Euprymna scolopes*. *Appl. Environ. Microbiol.* 78, 4200–4208. doi: 10.1128/AEM.07437-11

Collins, A., and Nyholm, S. (2011). Draft genome of *Phaeobacter gallaeciensis* ANG1, a dominant member of the accessory nidamental gland of *Euprymna scolopes*. *J. Bacteriol.* 193, 3397–3398. doi: 10.1128/JB.05139-11

Colston, S. M., Fullmer, M. S., Beka, L., Lamy, B., Gogarten, J. P., and Graf, J. (2014). Bioinformatic genome comparisons for taxonomic and phylogenetic assignments using *Aeromonas* as a test case. *MBio* 5:e02136. doi: 10.1128/mBio.02136-14

Cotter, P. D., Ross, R. P., and Hill, C. (2013). Bacteriocins – a viable alternative to antibiotics? *Nat. Rev. Microbiol.* 11, 95–105. doi: 10.1038/nrmicro2937

Csáky, T. Z. (1948). On the estimation of bound hydroxylamine in biological materials. *Acta Chem. Scand.* 2, 450–454. doi: 10.3891/acta.chem.scand.02-0450

Cude, W. N., Mooney, J., Tavanaei, A. A., Hadden, M. K., Frank, A. M., Gulvik, C. A., et al. (2012). Production of the antimicrobial secondary metabolite indigoidine contributes to competitive surface colonization by the marine

Roseobacter Phaeobacter sp. strain Y4I. *Appl. Environ. Microbiol.* 78, 4771–4780. doi: 10.1128/AEM.00297-12

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* 9:772. doi: 10.1038/nmeth.2109

Dogs, M., Teshima, H., Petersen, J., Fiebig, A., Chertkov, O., Dalingault, H., et al. (2013). Genome sequence of *Phaeobacter daeponensis* type strain (DSM 23529(T)), a facultatively anaerobic bacterium isolated from marine sediment, and emendation of *Phaeobacter daeponensis*. *Stand. Genomic Sci.* 9, 142–159. doi: 10.4056/sigs.4287962

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Fernandes, N., Case, R. J., Longford, S. R., Seyedsayamdost, M. R., Steinberg, P. D., Kjelleberg, S., et al. (2011). Genomes and virulence factors of novel bacterial pathogens causing bleaching disease in the marine red alga *Delisea pulchra*. *PLoS ONE* 6:e27387. doi: 10.1371/journal.pone.0027387

Fuchs, G. (1999). "Oxidation of organic compounds," in *Biology of the Prokaryotes*, eds J. W. Lengeler, G. Drews, and H. G. Schlegel (Stuttgart: Georg Thieme Verlag), 187–233.

Fullmer, M. S., Soucy, S. M., Swithers, K. S., Makkay, A. M., Wheeler, R., Ventosa, A., et al. (2014). Population and genomic analysis of the genus *Halorubrum*. *Front. Microbiol.* 5:140. doi: 10.3389/fmicb.2014.00140

Geng, H., Bruhn, J. B., Nielsen, K. F., Gram, L., and Belas, R. (2008). Genetic dissection of tropodithietic acid biosynthesis by marine roseobacters. *Appl. Environ. Microbiol.* 74, 1535–1545. doi: 10.1128/AEM.02339-07

Gibbs, K. A., Urbanowski, M. L., and Greenberg, E. P. (2008). Genetic determinants of self identity and social recognition in bacteria. *Science* 321, 256–259. doi: 10.1126/science.1160033

Gil-Turnes, M. S., Hay, M. E., and Fenical, W. (1989). Symbiotic marine bacteria chemically defend crustacean embryos from a pathogenic fungus. *Science* 246, 116–118. doi: 10.1126/science.2781297

Goetz, D. H., Holmes, M. A., Borregaard, N., Bluhm, M. E., Raymond, K. N., and Strong, R. K. (2002). The neutrophil lipocalin NGAL is a bacteriostatic agent that interferes with siderophore-mediated iron acquisition. *Mol. Cell* 10, 1033–1043. doi: 10.1016/S1097-2765(02)00708-6

González, J. M., Simó, R., Massana, R., Covert, J. S., Casamayor, E. O., Pedrós-Alió, C., et al. (2000). Bacterial community structure associated with a dimethylsulfoniopropionate-producing North Atlantic algal bloom. *Appl. Environ. Microbiol.* 66, 4237–4246. doi: 10.1128/AEM.66.10.4237-4246.2000

Grigioni, S., Boucher-Rodoni, R., Demarta, A., Tonolla, M., and Peduzzi, R. (2000). Phylogenetic characterisation of bacterial symbionts in the accessory nidamental glands of the sepioid *Sepia officinalis* (Cephalopoda: Decapoda). *Mar. Biol.* 136, 217–222. doi: 10.1007/s002270050679

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Jurado, R. L. (1997). Iron, infections, and anemia of inflammation. *Clin. Infect. Dis.* 25, 888–895. doi: 10.1086/515549

Kaufman, M., Ikeda, Y., Patton, C., van Dykhuizen, G., and Epel, D. (1998). Bacterial symbionts colonize the accessory nidamental gland of the squid Loligo opalescens via horizontal transmission. *Biol. Bull.* 194, 36–43. doi: 10.2307/1542511

Konstantinidis, K. T., Ramette, A., and Tiedje, J. M. (2006). Toward a more robust assessment of intraspecies diversity, using fewer genetic markers. *Appl. Environ. Microbiol.* 72, 7286–7293. doi: 10.1128/AEM.01398-06

Luo, H., Löytynoja, A., and Moran, M. A. (2012). Genome content of uncultivated marine *Roseobacters* in the surface ocean. *Environ. Microbiol.* 14, 41–51. doi: 10.1111/j.1462-2920.2011.02528.x

Luo, H., and Moran, M. A. (2014). Evolutionary ecology of the marine *Roseobacter* clade. *Microbiol. Mol. Biol. Rev.* 78, 573–587. doi: 10.1128/MMBR.00020-14

Luo, H., Swan, B. K., Stepanauskas, R., Hughes, A. L., and Moran, M. A. (2014). Evolutionary analysis of a streamlined lineage of surface ocean *Roseobacters*. *ISME J.* 8, 1428–1439. doi: 10.1038/ismej.2013.248

Martens, T., Heidorn, T., Pukall, R., Simon, M., Tindall, B. J., and Brinkhoff, T. (2006). Reclassification of *Roseobacter* gallaeciensis Ruiz-Ponte et al. 1998 as *Phaeobacter gallaeciensis* gen. nov., comb. nov., description of *Phaeobacter inhibens* sp. nov., reclassification of *Ruegeria algicola* (Lafay et al. 1995) Uchino et al. 1999 as Marinovu. *Int. J. Syst. Evol. Microbiol.* 56(pt 6), 1293–1304. doi: 10.1099/ijs.0.63724-0

McMillan, D. G. G., Velasquez, I., Nunn, B. L., Goodlett, D. R., Hunter, K. A., Lamont, I., et al. (2010). Acquisition of iron by alkaliphilic bacillus species. *Appl. Environ. Microbiol.* 76, 6955–6961. doi: 10.1128/AEM.01393-10

Miyashiro, T., and Ruby, E. G. (2012). Shedding light on bioluminescence regulation in *Vibrio fischeri*. *Mol. Microbiol.* 84, 795–806. doi: 10.1111/j.1365-2958.2012.08065.x

Moran, M. A., Buchan, A., González, J. M., Heidelberg, J. F., Whitman, W. B., Kiene, R. P., et al. (2004). Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. *Nature* 432, 910–913. doi: 10.1038/nature03170

Murdoch, S. L., Trunk, K., English, G., Fritsch, M. J., Pourkarimi, E., and Coulthurst, S. J. (2011). The opportunistic pathogen *Serratia marcescens* utilizes type VI secretion to target bacterial competitors. *J. Bacteriol.* 193, 6057–6069. doi: 10.1128/JB.05671-11

Newton, R. J., Griffin, L. E., Bowles, K. M., Meile, C., Gifford, S., Givens, C. E., et al. (2010). Genome characteristics of a generalist marine bacterial lineage. *ISME J.* 4, 784–798. doi: 10.1038/ismej.2009.150

Nyholm, S., Stewart, J., Ruby, E., and McFall-Ngai, M. (2009). Recognition between symbiotic *Vibrio fischeri* and the haemocytes of *Euprymna scolopes*. *Environ. Microbiol.* 11, 483–493. doi: 10.1111/j.1462-2920.2008.01788.x

Ong, D. S. T., Wang, L., Zhu, Y., Ho, B., and Ding, J. L. (2005). The response of ferritin to LPS and acute phase of *Pseudomonas* infection. *J. Endotoxin Res.* 11, 267–280. doi: 10.1179/096805105X58698

Ong, S. T., Ho, J. Z. S., Ho, B., and Ding, J. L. (2006). Iron-withholding strategy in innate immunity. *Immunobiology* 211, 295–314. doi: 10.1016/j.imbio.2006.02.004

Pichon, D., Gaia, V., Norman, M. D., and Boucher-Rodoni, R. (2005). Phylogenetic diversity of epibiotic bacteria in the accessory nidamental glands of squids (Cephalopoda: Loliginidae and Idiosepiidae). *Mar. Biol.* 147, 1323–1332. doi: 10.1007/s00227-005-0014-5

Pukatzki, S., Ma, A. T., Sturtevant, D., Krastins, B., Sarracino, D., Nelson, W. C., et al. (2006). Identification of a conserved bacterial protein secretion system in *Vibrio cholerae* using the *Dictyostelium* host model system. *Proc. Natl. Acad. Sci. U.S.A.* 103, 1528–1533. doi: 10.1073/pnas.0510322103

Rao, D., Webb, J. S., Holmström, C., Case, R., Low, A., Steinberg, P., et al. (2007). Low densities of epiphytic bacteria from the marine alga *Ulva australis* inhibit settlement of fouling organisms. *Appl. Environ. Microbiol.* 73, 7844–7852. doi: 10.1128/AEM.01543-07

Ravn, L., Christensen, A. B., Molin, S., Givskov, M., and Gram, L. (2001). Methods for detecting acylated homoserine lactones produced by Gram-negative bacteria and their application in studies of AHL-production kinetics. *J. Microbiol. Methods* 44, 239–251. doi: 10.1016/S0167-7012(01)00217-2

Raymond, K. N., Dertz, E. A., and Kim, S. S. (2003). Enterobactin: an archetype for microbial iron transport. *Proc. Natl. Acad. Sci. U.S.A.* 100, 3584–3588. doi: 10.1073/pnas.0630018100

Reasoner, D. J., and Geldreich, E. E. (1985). A new medium for the enumeration and subculture of bacteria from potable water. *Appl. Environ. Microbiol.* 49, 1–7. doi: 10.3891/acta.chem.scand.02-0450

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends Genet.* 16, 276–277. doi: 10.1016/S0168-9525(00)02024-2

Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U.S.A.* 106, 19126–19131. doi: 10.1073/pnas.0906412106

Riedel, T., Teshima, H., Petersen, J., Fiebig, A., Davenport, K., Daligault, H., et al. (2013). Genome sequence of the *Leisingera aquimarina* type strain (DSM 24565(T)), a member of the marine *Roseobacter* clade rich in extrachromosomal elements. *Stand. Genomic Sci.* 8, 389–402. doi: 10.4056/sigs.3858183

Rosemeyer, V., Michiels, J., Verreth, C., and Vanderleyden, J. (1998). luxI- and luxR-homologous genes of *Rhizobium etli* CNPAF512 contribute to synthesis of autoinducer molecules and nodulation of *Phaseolus vulgaris*. *J. Bacteriol.* 180, 815–821. doi: 10.1093/nar/gkh340

Ruiz-Ponte, C., Cilia, V., Lambert, C., and Nicolas, J. L. (1998). *Roseobacter gallaeciensis* sp. nov., a new marine bacterium isolated from rearings and collectors of the scallop *Pecten maximus*. *Int. J. Syst. Bacteriol.* 48(Pt 2), 537–542. doi: 10.1371/journal.pone.0027387

Russell, A. B., Hood, R. D., Bui, N. K., LeRoux, M., Vollmer, W., and Mougous, J. D. (2011). Type VI secretion delivers bacteriolytic effectors to target cells. *Nature* 475, 343–347. doi: 10.1038/nature10244

Salichos, L., and Rokas, A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497, 327–331. doi: 10.1038/nature12130

Schleicher, T., and Nyholm, S. (2011). Characterizing the host and symbiont proteomes in the association between the Bobtail squid, *Euprymna scolopes*, and the bacterium, *Vibrio fischeri*. *PLoS ONE* 6:e25649. doi: 10.1371/journal.pone.0025649

Schwyn, B., and Neilands, J. B. (1987). Universal chemical assay for the detection and determination of siderophores. *Anal. Biochem.* 160, 47–56. doi: 10.1126/science.1160033

Seyedsayamdost, M., Case, R., Kolter, R., and Clardy, J. (2011). The Jekyll-and-Hyde chemistry of *Phaeobacter gallaeciensis*. *Nat. Chem.* 3, 331–335. doi: 10.1126/science.2781297

Shikuma, N. I., Pilhofer, M., Weiss, G. L., Hadfield, M. G., Jensen, G. J., and Newman, D. K. (2014). Marine tubeworm metamorphosis induced by arrays of bacterial phage tail-like structures. *Science* 343, 529–533. doi: 10.1126/science.1246794

Soucy, S. M., Fullmer, M. S., Papke, R. T., and Gogarten, J. P. (2014). Inteins as indicators of gene flow in the halobacteria. *Front. Microbiol.* 5:299. doi: 10.3389/fmicb.2014.00299

Thole, S., Kalhoefer, D., Voget, S., Berger, M., Engelhardt, T., Liesegang, H., et al. (2012). *Phaeobacter gallaeciensis* genomes from globally opposite locations reveal high similarity of adaptation to surface life. *ISME J.* 6, 2229–2244. doi: 10.1007/s002270050679

van Heel, A. J., de Jong, A., Montalbán-López, M., Kok, J., and Kuipers, O. P. (2013). BAGEL3: automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. *Nucleic Acids Res.* 41, W448–W453. doi: 10.1093/nar/gkt391

Wagner-Döbler, I., Ballhausen, B., Berger, M., Brinkhoff, T., Buchholz, I., Bunk, B., et al. (2010). The complete genome sequence of the algal symbiont *Dinoroseobacter shibae*: a hitchhiker's guide to life in the sea. *ISME J.* 4, 61–77. doi: 10.1038/ismej.2009.94

Wagner-Döbler, I., and Biebl, H. (2006). Environmental biology of the marine *Roseobacter* lineage. *Annu. Rev. Microbiol.* 60, 255–280. doi: 10.1146/annurev.micro.60.080805.142115

Wang, D., Kim, B. Y., Lee, K. S., Yoon, H. J., Cui, Z., Lu, W., et al. (2009). Molecular characterization of iron binding proteins, transferrin and ferritin heavy chain subunit, from the bumblebee *Bombus ignitus*. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* 152, 20–27. doi: 10.1016/j.cbpb.2008.09.082

Whistler, C. A., and Ruby, E. G. (2003). GacA regulates symbiotic colonization traits of *Vibrio fischeri* and facilitates a beneficial association with an animal host. *J. Bacteriol.* 185, 7202–7212. doi: 10.1128/JB.185.24.7202

Wu, H.-Y., Chung, P.-C., Shih, H.-W., Wen, S.-R., and Lai, E.-M. (2008). Secretome analysis uncovers an Hcp-family protein secreted via a type VI secretion system in *Agrobacterium tumefaciens*. *J. Bacteriol.* 190, 2841–2850. doi: 10.1128/JB.01775-07

Zan, J., Cicirelli, E. M., Mohamed, N. M., Sibhatu, H., Kroll, S., Choi, O., et al. (2012). A complex LuxR-LuxI type quorum sensing network in a roseobacterial marine sponge symbiont activates flagellar motility and inhibits biofilm formation. *Mol. Microbiol.* 85, 916–933. doi: 10.1111/j.1365-2958.2012.08149.x

Zan, J., Liu, Y., Fuqua, C., and Hill, R. T. (2014). Acyl-homoserine lactone quorum sensing in the *Roseobacter* clade. *Int. J. Mol. Sci.* 15, 654–669. doi: 10.3390/ijms15010654

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 November 2014; accepted: 01 February 2015; published online: 23 February 2015.

*Citation: Collins AJ, Fullmer MS, Gogarten JP and Nyholm SV (2015) Comparative genomics of Roseobacter clade bacteria isolated from the accessory nidamental gland of Euprymna scolopes. Front. Microbiol. **6**:123. doi: 10.3389/fmicb.2015.00123*

This article was submitted to Microbial Symbioses, a section of the journal Frontiers in Microbiology.

Appendix D – Gromek, Suria et al., 2015

Leisingera sp. JC1, a Bacterial Isolate from Hawaiian Bobtail Squid Eggs, Produces Indigoidine and Differentially Inhibits Vibrios

This section sees my collaboration with the Nyholm lab continue. Andrea's paper is concerned with the repertoire of metabolic products here isolate from a squid egg jelly coat can produce (Gromek et al., 2016). My role was a reprisal of what I contributed to Collins et al., 2015 with the addition of some genome comparisons using BRIG and Mauve. I participated in the drafting of the relevant areas of the manuscript as well as in the editing of the document.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 08 September 2016
doi: 10.3389/fmicb.2016.01342

Leisingera sp. JC1, a Bacterial Isolate from Hawaiian Bobtail Squid Eggs, Produces Indigoidine and Differentially Inhibits Vibrios

Samantha M. Gromek[1†], Andrea M. Suria[2†], Matthew S. Fullmer[2], Jillian L. Garcia[1], Johann Peter Gogarten[2,3], Spencer V. Nyholm[2] and Marcy J. Balunas[1]**

[1] Division of Medicinal Chemistry, Department of Pharmaceutical Sciences, University of Connecticut, Storrs, CT, USA, [2] Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA, [3] Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA

Female members of many cephalopod species house a bacterial consortium in the accessory nidamental gland (ANG), part of the reproductive system. These bacteria are deposited into eggs that are then laid in the environment where they must develop unprotected from predation, pathogens, and fouling. In this study, we characterized the genome and secondary metabolite production of *Leisingera* sp. JC1, a member of the roseobacter clade (*Rhodobacteraceae*) of *Alphaproteobacteria* isolated from the jelly coat of eggs from the Hawaiian bobtail squid, *Euprymna scolopes*. Whole genome sequencing and MLSA analysis revealed that *Leisingera* sp. JC1 falls within a group of roseobacters associated with squid ANGs. Genome and biochemical analyses revealed the potential for and production of a number of secondary metabolites, including siderophores and acyl-homoserine lactones involved with quorum sensing. The complete biosynthetic gene cluster for the pigment indigoidine was detected in the genome and mass spectrometry confirmed the production of this compound. Furthermore, we investigated the production of indigoidine under co-culture conditions with *Vibrio fischeri*, the light organ symbiont of *E. scolopes*, and with other vibrios. Finally, both *Leisingera* sp. JC1 and secondary metabolite extracts of this strain had differential antimicrobial activity against a number of marine vibrios, suggesting that *Leisingera* sp. JC1 may play a role in host defense against other marine bacteria either in the eggs and/or ANG. These data also suggest that indigoidine may be partially, but not wholly, responsible for the antimicrobial activity of this squid-associated bacterium.

Keywords: symbiosis, *Euprymna*, roseobacter, *Rhodobacteraceae*, indigoidine, *Leisingera*, DART-MS, secondary metabolite regulation

OPEN ACCESS

Edited by:
Russell T. Hill,
University of Maryland Center for Environmental Science, USA

Reviewed by:
Valerie McKenzie,
University of Colorado Boulder, USA
Anahit Penesyan,
Macquarie University, Australia

****Correspondence:***
Marcy J. Balunas
marcy.balunas@uconn.edu
Spencer V. Nyholm
spencer.nyholm@uconn.edu

†These authors have contributed equally to this work.

Specialty section:
This article was submitted to Microbial Symbioses, a section of the journal Frontiers in Microbiology

Received: *23 May 2016*
Accepted: *15 August 2016*
Published: *08 September 2016*

Citation:
Gromek SM, Suria AM, Fullmer MS, Garcia JL, Gogarten JP, Nyholm SV and Balunas MJ (2016) Leisingera sp. JC1, a Bacterial Isolate from Hawaiian Bobtail Squid Eggs, Produces Indigoidine and Differentially Inhibits Vibrios. Front. Microbiol. 7:1342. doi: 10.3389/fmicb.2016.01342

INTRODUCTION

It is becoming increasingly evident that many animals and plants use compounds produced by symbiotic bacteria for protection against pathogens and other fouling organisms (reviewed in Flórez et al., 2015). In marine and aquatic environments a number of invertebrates (including sponges, tunicates, bryozoans, and molluscs) host microorganisms that produce compounds used for such protection. These groups have served as an important source for studying defensive symbioses and for the discovery of novel bioactive natural products (see example in Schmidt and Donia, 2010).

Among molluscs, one common yet poorly understood animal–bacterial association occurs between members of squid and cuttlefish species and bacterial consortia that reside within a reproductive gland of female hosts called the accessory nidamental gland (ANG; Kaufman et al., 1998; Grigioni et al., 2000; Barbieri et al., 2001; Pichon et al., 2005; Collins et al., 2012). This organ harbors a dense consortium of bacteria housed in epithelium-lined tubules that are attached to the nidamental gland, the organ that secretes the jelly coat (JC) surrounding fertilized eggs. Bacteria from the ANG are deposited into the JC where they have been hypothesized to help protect developing eggs from fouling microorganisms, pathogens, and/or predation (Barbieri et al., 1997, 2001; Collins et al., 2012, 2015).

The Hawaiian bobtail squid, *Euprymna scolopes*, has been used as a model organism to study bacteria–host interactions, mainly due to the host's relationship with the bioluminescent bacterium *Vibrio fischeri* (McFall-Ngai, 2014). Recent studies have also focused on a second association found within the ANG of this species (Collins and Nyholm, 2011; Collins et al., 2012, 2015). These studies demonstrated that the ANG consortium in *E. scolopes* is dominated by members of the *Rhodobacteraceae* (roseobacters) within the *Alphaproteobacteria*, a common group of marine bacteria. A number of roseobacter-clade organisms are known to produce unique antimicrobial molecules and other secondary metabolites. For example, the antibiotic tropodithietic acid (TDA) and the algicidal roseobacticides are produced by *Phaeobacter* species and the antibacterial compound indigoidine is produced by *Leisingera* (formerly *Phaeobacter*) sp. Y4I (Geng et al., 2008; Seyedsayamdost et al., 2011; Cude et al., 2012). Most of these studies have focused on either free-living or plankton-associated roseobacters and the potential antimicrobial activity of the ANG strains has not been explored. A study that analyzed the genomes of 13 ANG roseobacter strains from *E. scolopes* did reveal the potential for secondary metabolite production (Collins et al., 2015) and *Gammaproteobacteria* from the ANG of another squid species have been shown to inhibit other bacteria (Barbieri et al., 1997).

In this study, we characterized the genome and secondary metabolite production of a new bacterial strain, *Leisingera* sp. JC1, isolated from the JC of *E. scolopes* squid eggs. Whole genome sequencing and biochemical analyses revealed the potential for and production of a number of secondary metabolites, including siderophores and acyl-homoserine lactones involved with quorum sensing. The complete indigoidine biosynthetic gene cluster was detected in the genome and mass spectrometry confirmed the production of this compound. Furthermore, we investigated the regulation of indigoidine under co-culture conditions with *V. fischeri*, the light organ symbiont. Finally, both *Leisingera* sp. JC1 and extracts from this strain exhibited differential antimicrobial activity against a number of marine vibrios, suggesting that indigoidine may be partially, but not wholly, responsible for the antimicrobial activity of this squid-associated bacterium.

MATERIALS AND METHODS

Bacterial Isolation

Hawaiian bobtail squid, *E. scolopes*, were obtained from sand flats in Oahu (Maunalua Bay, 21°16′51.42″ N, 157°43′33.07″ W), Hawaii and maintained in aquaria as previously described (Schleicher and Nyholm, 2011). Eggs laid in captivity from one adult female were collected, flash frozen on the 11th day of development, and stored at −80°C. Ten eggs were thawed for bacterial isolation and their outer capsules and embryos were removed and discarded with sterile forceps. The JCs were isolated, surface sterilized with 70% ethanol, and rinsed with filter-sterilized squid Ringers (FSSR, 530 mM NaCl, 25 mM $MgCl_2$, 10 mM $CaCl_2$, 20 mM HEPES, pH = 7.5). The 10 JCs were pooled and homogenized in FSSR, then serially diluted and plated on seawater tryptone (SWT) medium (5 g/L tryptone, 3 g/L yeast extract, 3 mL/L glycerol, 700 mL/L Instant Ocean sea salts, 15 g/L agar, 300 mL/L DI water). *Leisingera* sp. JC1 colonies appeared dark blue on this medium and were streaked to isolation.

Genomic Sequencing and Analysis

Genomic DNA was extracted using the MasterPure DNA Purification kit (Epicentre, Madison, WI, USA) from an overnight liquid culture of *Leisingera* sp. JC1 grown shaking at 30°C in SWT. DNA was quantified using a Qubit 2.0 fluorometer (Life Technologies, Agawam, MA, USA) and checked for quality on a 1% agarose gel and using a NanoDrop 1000 spectrophotometer (Thermo Scientific, Agawam, MA, USA). A paired end library was prepared from 1 ng of genomic DNA using the Nextera XT DNA library kit (Illumina, Inc., San Diego, CA, USA) and quantified using the Qubit fluorometer and bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). The library was sequenced on an Illumina MiSeq sequencer using 2 bp × 250 bp reads at the Microbial Analysis Resources and Services (MARS) facility at the University of Connecticut (Storrs, CT, USA).

Reads were trimmed using the CLC Genomic Workbench (Qiagen, Hilden, Germany) and a draft genome was assembled using the A5 assembler (Tritt et al., 2012). Coverage was determined by mapping trimmed reads to the draft genome assembly using CLC Genomic Workbench. The genome was annotated using the Rapid Annotation using Subsystem Technology (RAST, Aziz et al., 2008)[1] server and analyzed with the Antibiotic and Secondary Metabolite Analysis Shell 3.0 (antiSMASH, Weber et al., 2015)[2] for potential secondary metabolite biosynthesis gene clusters. The draft genome assembly has been deposited in DDBJ/EMBL/GenBank under accession LYUZ00000000. The version described in this paper is version LYUZ01000000.

[1] http://rast.nmpdr.org

[2] http://antismash.secondarymetabolites.org

Taxonomic Analysis and Whole Genome Comparison

Initial 16S identity suggested JC1 belonged to the genus *Leisingera* (data not shown). To validate this conclusion and to evaluate its relationship to the previously sequenced ANG isolates, a further taxonomic analysis was undertaken that used 17 previously described *Leisingera* genomes (Collins et al., 2015). A 33 gene multilocus sequence analysis was carried out following the methodology described in Collins et al. (2015). After generating alignments for each of the 33 genes using MUSCLE (Edgar, 2004), a concatenated alignment was generated using in-house python scripts. An optimal model of evolution was determined using the Akaike information criterion with correction for small sample size as implemented in jModelTest v2.1.4 (Darriba et al., 2012). The best-fitting model reported was GTR + Gamma estimation + Invariable site estimation. A maximum-likelihood (ML) phylogeny was generated from the concatenated multi-sequence alignment using PhyML v3.0_360-500M (Guindon et al., 2010). PhyML parameters consisted of GTR model, estimated p-invar, four substitution rate categories, estimated gamma distribution, sub-tree pruning and regrafting enabled with 100 bootstrap replicates. In addition to the maximum-likelihood analysis, a Bayesian inference analysis was also conducted using MrBayes v3.2.4 x64 (Ronquist et al., 2012). A mixed model with gamma estimation and invariable sites was used. The mixed model settled on a GTR submodel with only one parameter difference from the default GTR model with a posterior probability > 0.8. The standard GTR model accounted for the remainder of the model probability. The analysis used two cold chains with three heated chains each and ran for one million generations. After the run finished, convergence was assessed using average standard deviation of split frequencies of the cold chains, potential scale reduction factors of parameters, and minimum effective sample sizes of parameters. All criteria indicated the runs had converged.

Average nucleotide identity (ANI) was calculated using JSpecies 1.2.1 (Richter and Rosselló-Móra, 2009). The calculations were made using the MUMmer aligner with its default options. Contig files were generated for this analysis using the seqret function of the EMBOSS package (Rice et al., 2000). The reciprocal comparisons were averaged for reporting. Estimates of *in silico* DDH were made using the Genome-to-genome distance calculator 2.1 (Meier-Kolthoff et al., 2013) using the BLAST+ alignment method and the formula 2 algorithm outputs.

Select genomes were compared using the BLAST Ring Generator (BRIG) v1.0 (Alikhan et al., 2011). Default BLAST options were used. A whole genome alignment was generated using the Mauve program v2.3.1 (Darling et al., 2010). The progressiveMauve algorithm was used with default options.

Homoserine Lactone Detection

Homoserine lactone (HSL) production was detected using a well-diffusion assay with the HSL-sensing bacterium *Agrobacterium tumefaciens* NTL4 pZLR4 (Cha et al., 1998) as previously described (Ravn et al., 2001; Collins et al., 2015). In brief, *A. tumefaciens* NTL4 was grown in 3 mL of LB with 30 μg/mL gentamicin for 24 h at 30°C. This culture was used to inoculate 50 mL of AB minimal media with 0.5% casamino acids and 0.5% glucose (Chilton et al., 1974), and allowed to grow for another 24 h at 30°C. This culture was used to inoculate 100 mL of AB minimal media to which 1.2% agar had been added and a final concentration of 0.5% casamino acids, 0.5% glucose, and 75 μg/mL 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) was added after autoclaving. The inoculated molten agar was allowed to solidify in Petri dishes and wells were cut into the media using a sterile borer.

Leisingera sp. JC1 and *Leisingera* sp. ANG1 were grown overnight at 30°C in 3 mL of SWT broth with 30 μM $FeCl_3$, 0.5% glucose, and 0.5% casamino acids to induce HSL production. Cells were pelleted and the supernatant was collected and filtered through a 0.22 μm filter (Thermo Scientific, Agawam, MA, USA). Cell-free supernatant (60 μl) was added to the wells in the *A. tumefaciens* plates. The *N*-3-oxohexanoyl homoserine lactone standard was serially diluted and added to wells of an *A. tumefaciens* plate as a control and for semi-quantitative comparison. All plates were incubated at 28°C for 24 h before imaging.

Siderophore Detection

To qualitatively detect siderophore production, *Leisingera* sp. JC1 was plated in triplicate on chrome azurol S (CAS) indicator agar, modified for marine bacteria as previously described (Whistler and Ruby, 2003), and incubated at 28°C for 24 h before imaging. Sequestration of iron from CAS causes a color change from blue to orange, indicating siderophore production.

Detection of Indigoidine Biosynthesis Genes in JC1 Genomic DNA

To confirm the presence of indigoidine biosynthesis genes in JC1, genomic DNA was extracted and quantified as described for genomic sequencing above. Primers were designed (Supplemental Table S1) to amplify the *igiCDR* genes based on the draft genome assembly and using Primer3 software (Untergasser et al., 2012). PCR amplification was performed using the standard GoTaq Green Master Mix (Promega, Madison, WI, USA) protocol with 30 cycles and 55°C annealing temperature.

Leisingera sp. JC1 Large Scale Culture

Leisingera sp. JC1 was cultured for extraction using SWT media (as described above except without addition of glycerol, delineated hereafter as SWT_{ng}). A three step culturing process was employed to produce sufficient scale for secondary metabolite extraction, while ensuring that the bacterium was in late stationary phase for optimal production of secondary metabolites (Ruiz et al., 2010). First, small scale cultures were prepared by inoculating a JC1 colony into 5 mL of media in a 24 deep well plate, which was incubated for 3 days at room temperature while shaking at 200 rpm. Then, medium scale cultures were prepared by transferring 1.5 mL of the small scale cultures into 125 mL baffled flasks with 50 mL media, which were incubated for 3 days at room temperature while shaking at 125 rpm. Lastly, large scale cultures were prepared by transferring

15 mL of medium scale cultures into 1 L baffled flasks with 500 mL of media, which were incubated for 3 days at room temperature while shaking at 125 rpm.

Extraction of *Leisingera* sp. JC1

All extraction solvents were ACS grade and purchased from Sigma Aldrich (St. Louis, MO, USA).

Normal Extraction

Diaion HP20 resin (Supelco, Bellefonte, PA, USA) was pre-washed by sequentially rinsing resin with methanol and Millipore water (EMD Millipore, Billerica, MA, USA). Large scale JC1 cultures were sonicated to lyse cells prior to addition of pre-washed Diaion HP20 resin (50 g, 10% w/v), followed by incubation for 24 h at room temperature while shaking at 125 rpm. Bacterial culture and resin were then filtered using a coarse glass frit filter and washed with Millipore water to remove aqueous media components. The resin and bacterial culture were then sequentially extracted with methanol, dichloromethane, and acetone (2 × 150 mL). Organic portions were combined, extracted with ethyl acetate to remove residual aqueous material, and concentrated.

Indigoidine Enriched Extraction

Because indigoidine is poorly soluble in water and most organic solvents, a second extraction protocol was utilized to prepare an indigoidine enriched extract following modified literature procedures (Yu et al., 2013). Briefly, large scale cultures were sonicated to lyse cells and transferred to centrifuge tubes. Cells were then separated from supernatant by low-speed centrifugation (850 *g* × 5 min; Beckman Coulter Avanti J-E Centrifuge, Brea, CA, USA). Supernatant was transferred to new tubes and subjected to high-speed centrifugation (21,000 *g* × 10 min) to obtain an indigoidine enriched pellet. The pellet was washed with methanol, transferred to a microcentrifuge tube, dried under N_2 gas, and dissolved in dimethyl sulfoxide (DMSO).

Detection of Indigoidine Production by *Leisingera* sp. JC1 via LC–MS

All HPLC grade solvents and reagents were purchased from Sigma-Aldrich. LC–MS data were collected on an Agilent ESI single quadrupole mass spectrometer coupled to an Agilent 1260 HPLC system with a G1311 quaternary pump, G1322 degasser, and a G1315 diode array detector (Agilent Technologies, Santa Clara, CA). A gradient elution was used from 10% methanol in H_2O to 90% methanol in H_2O over 25 min using an Agilent Eclipse XDB-C_{18} RP-HPLC column (4.6 mm × 150 mm, 5 μm) and a flow rate of 1 mL/min. Indigoidine enriched extracts were prepared at 5 mg/mL in DMSO. Indigoidine eluted at retention time (t_R) 10.7 min in agreement with literature (Yu et al., 2013).

Zone of Inhibition Assays

To observe inhibition of vibrio strains and ANG isolate strains by *Leisingera* sp. JC1 (Supplementary Table S3), a zone of inhibition (ZOI) assay was used. The vibrio strains *V. anguillarum* 775, *V. parahaemolyticus* KNH1, *V. fischeri* ES114, *V. harveyi* B392, and *Photobacterium leiognathi* KNH6 were grown for 2.5 h (to stationary phase) at 30°C in YTSS (4 g/L tryptone, 2.5 g/L yeast extract, 15 g/L Instant Ocean sea salts) broth and then serially diluted from 10^7 to 10^4 CFU/mL in YTSS broth to observe density dependent inhibition. Each dilution was plated in triplicate on YTSS agar using a sterile swab to form a lawn. All ANG isolates tested were grown overnight (~4×10^8 CFU/mL) in SWT broth at 30°C and plated on SWT agar using a sterile swab to form a lawn. *Leisingera* sp. JC1 was grown overnight to a density of ~1×10^8 CFU/mL in SWT when testing with ANG isolates and in YTSS when testing with vibrio strains. This overnight broth of *Leisingera* sp. JC1 was spotted (10 μL) on the surface of each lawn in quadruplicate. All plates were incubated at 28°C for 24 h before imaging and ZOI measurements around the *Leisingera* sp. JC1 colonies. SWT or YTSS broth (10 μL) were spotted on each lawn as media controls, and 10 μL of the overnight culture of *Leisingera* sp. JC1 was spotted in quadruplicate on SWT or YTSS agar without any bacterial lawns as a growth control.

To quantify inhibition, an average of three ZOI diameters were measured and an average of three diameters of the JC1 colonies were measured using ImageJ (Schneider et al., 2012). Due to slight variations in JC1 colony size across trials, the measurements were normalized by subtracting the average JC1 colony diameter from the average ZOI diameter. To determine if differences in ZOIs across lawn densities per organism were statistically significant, one-way ANOVAs were performed. If the results of the one-way ANOVA indicated statistically significant differences, multiple comparisons *post hoc* Tukey tests were performed to determine which lawn densities were significantly different.

96-Well Liquid Assays

Leisingera sp. JC1 extracts were tested for antibacterial activity against *V. fischeri* ES114, *V. anguillarum* 775, and *V. parahaemolyticus* KNH1. High throughput assays with these bacterial strains were developed based on similar assays with natural product extracts and human pathogens (Zgoda and Porter, 2001), including obtaining CFU counts and growth curves for each of the vibrio strains as well as determining proper incubation times and temperatures and finding appropriate controls. These assays were performed in 96-well plates (Corning Costar, Corning, NY, USA) with SWT media and incubated at 28°C while shaking at 200 rpm. The bacterial inocula were prepared by adding select colonies into 5 mL of media and adjusted to OD_{600} 0.1 (approximately 1–2 × 10^8 CFU/mL as per Clinical and Laboratory Standards Institute, 2012). Colony forming unit (CFU) counts were manually confirmed to ensure accurate approximation for each vibrio strain.

Extracts were screened as previously described (Zgoda and Porter, 2001) with the following modifications. Briefly, master mix was prepared by addition of 1.6 mL adjusted vibrio inoculum, 7.84 mL sterile water, and 6.4 mL of SWT media. To each well, 198 μL of master mix was added with 2 μL of either positive control (chloramphenicol, final testing concentration 2.5 μg/mL), negative control (DMSO), or extract prepared in DMSO (screened at final concentration of 500 μg/mL; MIC performed using serial dilutions). Sterility control wells consisted

of 98 μL sterile water, 100 μL of SWT media, and 2 μL of DMSO. All controls and samples were tested in technical triplicates with experiments repeated a minimum of three times to confirm results. Plates were read at 600 nm every 2 h from 0 to 10 h with a final reading at 24 h using a Synergy H1 Hybrid Reader (Biotek, Winooski, VT, USA). Results are given as percent control activity (PCA) calculated in comparison with DMSO, the negative control.

Localization of Indigoidine Production by *Leisingera* sp. JC1 Using DART-MS

Direct analysis in real time-mass spectrometry (DART-MS) analysis was performed using a JEOL AccuTOF with DART ion source (IonSense, Inc., Saugus, MA, USA). High purity helium 5.0–6.0 grade (greater than 99.999% purity) was heated to 300°C and used for ionization. Five locations were selected on JC1 colonies in the presence or absence of *V. fischeri*, including (A) center of colony, (B) midpoint between center and edge of colony, (C) edge of colony, (D) ZOI (in the absence of *V. fischeri* sample was obtained from a point equidistant from colony edge), and (E) outside ZOI. At each location a sterile single use syringe needle (BD Medical, Franklin Lakes, NJ, USA) was placed in the sample and then placed between the DART ion source and the MS inlet. Positive ion MS data were obtained over a *m/z* range of 60–700 and relative percent abundance was obtained for the indigoidine ion. Standards were run after sampling each colony and mass spectral data were monitored in real time to ensure no residual indigoidine remained after each sample. DART-MS is only semi-quantitative due to the potential for differential ionization, suppression of ions, and/or changes in sample concentration in the DART ion source (Sanchez et al., 2011). Therefore, relative indigoidine ion abundance was used to generate heatmaps representing a gradient from less abundance (black) to more abundance (red).

Measurement of Indigoidine Production by *Leisingera* sp. JC1 in Co-culture

JC1 bacterial inoculum was prepared by adding JC1 colonies into 5 mL of SWT_{ng} media in a 24 deep well plate, incubated for 24 h at room temperature while shaking at 200 rpm. Bacterial inocula for the vibrios were prepared by adding bacterial colonies of each species separately into 5 mL of SWT_{ng} media in 24 deep well plates, incubated for 2 h at 28°C while shaking at 200 rpm. All bacterial inocula (JC1, *V. fischeri*, *V. anguillarum*, *V. parahaemolyticus*) were adjusted to OD_{600} 0.1 prior to use.

Co-cultures of JC1 with individual vibrios were prepared by adding 1 mL of adjusted JC1 inoculum to 10 mL SWT_{ng} media in 125 mL baffled flasks, incubated for 24 h at room temperature while shaking at 125 rpm, followed by addition of 200 μL of *V. fischeri*, *V. anguillarum*, or *V. parahaemolyticus*. After addition of the vibrio strain, co-cultures were incubated for an additional 24 h at room temperature while shaking at 125 rpm. Monocultures of JC1, *V. fischeri*, *V. anguillarum*, and *V. parahaemolyticus* were prepared by adding 1 mL of adjusted inoculum to 10 mL SWT_{ng} media in 125 mL baffled flasks, incubated for 48 h while shaking at 125 rpm.

All co-cultures and monocultures were extracted using the indigoidine enriched protocol described above. LC–MS data was obtained on the Agilent LC–MS system described above, using an isocratic method to ensure minimal baseline variation (10% acetonitrile in H_2O with 0.1% formic acid over 15 min at a flow rate of 1 mL/min with 20 μL injection volume). Extracts were prepared at 5 mg/mL in DMSO. Indigoidine was detected and quantitated via measurement of area under the curve at UV absorbance 299 nm and confirmed by MS.

RESULTS AND DISCUSSION

Genome Characteristics and General Metabolism

Taxonomic Placement of JC1

Leisingera sp. JC1 has a draft genome size of 5.19 Mb and GC content of 62.3% (**Table 1**), which is average for members of the roseobacter clade and similar to other squid-associated isolates (Collins et al., 2015). This larger genome size reflects the generalist lifestyle and ability to use diverse energy sources common of roseobacters (Newton et al., 2010). The *repABC* genes for plasmid replication are present as well as *tra* genes necessary for conjugative plasmid transfer, indicating the potential presence of extrachromosomal DNA. Further sequencing is necessary to confirm the number, size, and content of these putative plasmids.

Phylogenetic reconstruction methods (Bayesian and maximum-likelihood) used with the 33-gene concatenation returned identical topologies with overall strong statistical supports (**Figure 1**), placing JC1 close to the *Leisingera* taxa previously isolated from the ANG. Average nucleotide identity (ANI) and *in silico* DNA–DNA hybridization estimates (*is*DDH) support this placement. JC1 had higher ANI (90.5–91.7%) (Supplementary Table S2) and *is*DDH (38.8–44.6%) values with the ANG isolates than with any other *Leisingera* sp. These results also show that JC1 does not group with either *Leisingera* sp. ANG-M7 or the ANG1 group, but is still related to both (**Figure 1**; Supplementary Table S3), which is not unusual since other ANG isolates also fall outside the main ANG1 clade (Collins et al., 2015).

There are indications that JC1 may be more similar to *Leisingera* sp. ANG-M7 than to the *Leisingera* sp. ANG1 group. Both the ANI and *is*DDH values between JC1 and ANG-M7 are elevated in comparison to their values with the ANG1 group. There are no support statistics for ANI so it is uncertain if the 1.2% (JC1-M7 ANI versus JC1 compared to the ANG group) and 1.6% (M7-JC1 ANI versus M7 compared to the ANG group) higher values are significantly different. However, *is*DDH values are supported by 95% confidence intervals. The lower interval for JC1-M7 does not overlap with the upper interval for any comparison with a member of the ANG group, suggesting the *is*DDH values are significantly different. Additionally, the Bayesian inference found a small fraction of topologies in which the placements of JC1 and ANG-M7 were reversed, while the maximum-likelihood analysis found this occurrence in 33 of

TABLE 1 | Genome statistics of *Leisingera* sp. JC1.

Genome size (Mb)	Number of contigs	N_{50} (bp)	G + C content (%)	Number of genes	Missing genes* (% of total)	Fold coverage
5.19	168	123,213	62.3	5,074	54 (1.1)	37

**As predicted by the RAST server (Aziz et al., 2008).*

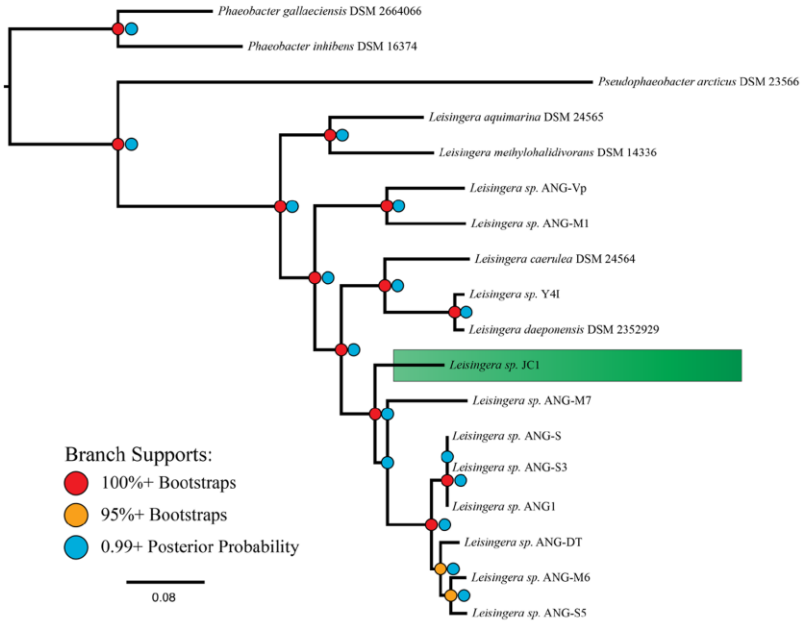


FIGURE 1 | MLSA analysis places *Leisingera* sp. JC1 in a sister group to other previously isolated ANG bacteria. Based on a comparison of 33 single-copy housekeeping genes, the egg jelly coat (JC) isolate JC1 is placed in the previously described roseobacter clade, "Clade 1" (Newton et al., 2010), along with seven previously isolated ANG bacteria (Collins et al., 2015). The phylogenetic tree reflects both Bayesian inference and maximum-likelihood methods with both posterior probability and bootstrap supports displayed.

100 bootstrap replicates. Overall, these analyses suggest that *Leisingera* sp. JC1 is distinct from, but related to the current ANG isolates.

Isolates having similar pigmentation to *Leisingera* sp. JC1 were cultured from other egg clutches, an ANG, and ovary from different females (data not shown). Among these, colonies with a similar dark blue morphology were isolated from the JCs of 1 and 23 day old eggs laid by different females. Similar colonies were isolated from the ANG of one of these females and the ovary of another female. Preliminary 16S sequencing placed two of these isolates in the genus *Leisingera* (data not shown), and further sequencing will reveal if these are the same strain as JC1. In addition, the production of the pigment indigoidine was confirmed by these additional strains (see below). These data suggest that *Leisingera* sp. JC1 and/or other indigoidine-producing strains may be selected for in the ANG/JC symbiosis.

Primary Metabolism

Leisingera sp. JC1 has a complete Entner-Duodoroff pathway and tricarboxylic acid cycle for metabolism of glucose. JC1 lacks any orthologs of phosphofructokinase, a major enzyme of glycolysis, but does contain a glucokinase and two distinct glucose-6-phosphate-1-dehydrogenases (GAPDHs). A glucose-6-phosphate-1-dehydrogenase (GPDH) is present, which catalyzes the first step of the alternative pathways for glucose metabolism, indicating that the Entner-Duodoroff pathway is probably used instead of glycolysis. *Leisingera* sp. JC1 only has the first two enzymes of the oxidative pentose phosphate pathway, but any 6-phosphate-gluconate produced can be further dehydrated by the Entner-Duodoroff pathway. Glycolate is a dissolved organic carbon often excreted by phytoplankton, and can be a carbon source for marine heterotrophic bacteria (Edenborn and Litchfield, 1985). *Leisingera* sp. JC1 is predicted

to oxidize glycolate to glyoxylate by a glycolate oxidase. JC1 has one system for glycerol uptake, the Ugp system, which can transport glycerol-3-phosphate against the concentration gradient. Sulfur oxidation genes are present, as well as a complete denitrification pathway with a copper-containing nitrite reductase. An assimilatory nitrate reductase is also present, which can convert nitrate to nitrite. An ammonia assimilation pathway is present with a ferrodoxin-dependent GOGAT, but no adenylyltransferase gene (GlnE) is present.

Transport

The high-affinity inorganic phosphate transport genes *pstABCS* and their regulatory genes *phoBUR* are present in JC1. The siderophore biosynthesis genes *asbAB* and *siderX456*, which encode high-affinity iron chelators, and the ferric iron ABC transporter, *pitADC*, are also present. JC1 has ABC transporters for dipeptides, oligopeptides, branched-chain amino acids, alkylphosphonate, and tungstate. The tripartite ATP-independent periplasmic (TRAP) transporter genes *dctMPQ* are present for unknown substrates, as well as the twin-arginine translocation (TAT) system genes, *tatABC*.

Leisingera sp. JC1 contains all 13 genes that encode the structural proteins essential for the Type VI Secretion System (T6SS) to function (Cianfanelli et al., 2016). The T6SS is a one-step mechanism for delivery of effectors across the Gram-negative outer membrane and membrane of the target cell, be it bacterial or eukaryotic. Widespread amongst the *Proteobacteria*, some T6SSs have been implicated in eukaryotic virulence (Pukatzki et al., 2006; Sana et al., 2012), but the majority are believed to play a role in bacterial competition (Hood et al., 2010; Schwarz et al., 2010). While it is possible for one T6SS system to affect both bacterial and eukaryotic targets (Jiang et al., 2014) it is believed that the system evolved for interactions with other bacteria, even in the case of intraspecific competition (Unterweger et al., 2014). Little work has been done, however, to investigate the role of T6SSs in beneficial host-symbiont relationships. Eleven of the 12 previously described ANG isolates also possess a T6SS (Collins et al., 2015), and it is possible that this system plays a role in interactions with other ANG or JC bacteria and/or the squid host. In the ANG, bacteria are partitioned into densely packed, epithelium-lined tubules, where each tubule is dominated by a particular taxon (Collins et al., 2012). These ANG/JC isolates may utilize the T6SS to outcompete other bacteria to establish colonization of a single tubule. While *Leisingera* sp. JC1 groups closely with other ANG isolates that also possess a T6SS (**Figure 1**, Collins et al., 2015), intraspecific effectors may facilitate competition between these strains, since ANG tubules are often highly pigmented with a single color (e.g., all dark blue matching the pigmentation of JC1 or all red-orange matching the pigmentation of several ANG isolates). Future studies will investigate the nature of JC1's T6SS effector proteins in the ANG symbiosis. There are numerous classes of evolved effector VgrG proteins, each with their own enzymatic function (reviewed in Durand et al., 2014). Understanding the number and type of effectors that JC1 can produce and deliver may help elucidate any role in the symbiosis.

Secondary Metabolite Biosynthesis

Analysis with the antibiotic and Secondary Metabolite Analysis Shell (antiSMASH, Weber et al., 2015) predicted several potential secondary metabolite biosynthesis gene clusters (Supplementary Table S4). These results included three separate siderophore clusters, one bacteriocin, one HSL, one type 1 polyketide synthase (T1 PKS), one other PKS (not type 1,2,3, or *trans*-AT), and two clusters classified as "other." Of these two "other" clusters, one contains the biosynthesis cluster for the known antimicrobial metabolite, indigoidine (Cude et al., 2012), while the other contains a previously described putative hybrid polyketide synthase/non-ribosomal peptide synthetase (PKS/NRPS) gene cluster known to be conserved amongst roseobacters (Martens et al., 2007). This PKS/NRPS gene cluster encodes a polyketide synthase, glycosyl transferase, non-ribosomal peptide synthetase, and phosphopantetheinyl transferase, but the product of this cluster has not yet been identified. The top homologous gene cluster of the T1 PKS is 45% similar to a cluster in the ANG isolate, *Leisingera* sp. ANG-M7. While some roseobacters are capable of producing the novel secondary metabolite TDA (Bruhn et al., 2006, 2007; Geng et al., 2008), genes for synthesis of this molecule were not found nor was the molecule detected via LC–MS (data not shown).

Quorum Sensing

AntiSMASH predicted one *luxIR* homolog in *Leisingera* sp. JC1, flanked by an acyltransferase, crotonyl-CoA reductase, helicase, and oxidoreductase, similar to the previously published gene arrangement in bacterial isolates from the ANG (Collins et al., 2015). Production of HSLs by JC1 was confirmed in the *A. tumefaciens* NTL4 reporter assay, in which cell-free supernatant of a JC1 culture did induce β-galactosidase activity, indicating the presence of HSLs (**Figures 2A,B**). When compared to a dilution series of the *N*-3-oxohexanoyl HSL, JC1 produced a halo similar to that seen by 25 nM of HSL standard. The HSL production of JC1 was also slightly less than that of a closely related ANG isolate, *Leisingera* sp. ANG1.

Understanding the gene regulation by quorum sensing will be an important avenue of research for *Leisingera* sp. JC1 and the other *E. scolopes* ANG isolates due to the different habitats these bacteria experience. It is hypothesized that cephalopod ANGs are colonized via horizontal transmission from the environment (Kaufman et al., 1998), and potential symbionts must switch from living at very low cell densities in the seawater to very high cell densities in the ANG tubules (Collins et al., 2012). When ANG bacteria are deposited into the JC layers of eggs, these bacteria again experience a switch from the very high densities of the ANG to a lower density in the eggs. Due to this change in environments and cell densities, quorum sensing may play a role in gene regulation for ANG/egg JC bacteria.

Quorum sensing is also important in host-microbe interactions involving other roseobacters. For example, quorum sensing regulates motility and biofilm formation during host colonization in the sponge symbiont *Ruegeria* sp. KLH11 (Zan et al., 2012) and is necessary for colonization of the alga, *Ulva*

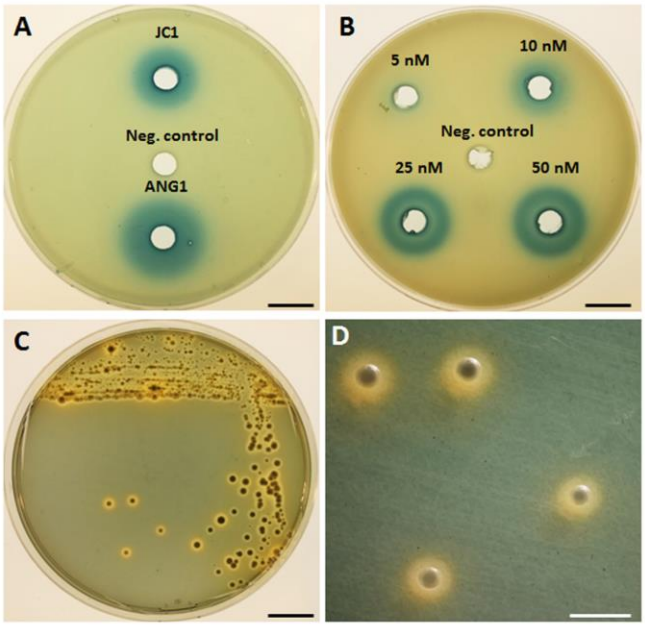


FIGURE 2 | Detection of homoserine lactone production and siderophore production by *Leisingera* sp. JC1. (A) Homoserine lactones were detected by β-galactosidase activity in cell-free supernatant of *Leisingera* sp. JC1 and compared with previously tested *Leisingera* sp. ANG1 (Collins et al., 2015); SWT broth was used as a negative control. **(B)** Dilution of *N*-3-oxohexanoyl homoserine lactone used as a positive control for the HSL assay; DMSO was the HSL standard solvent and was the negative control. **(C)** *Leisingera* sp. JC1 plated on CAS agar. Sequestration of iron changes the media from blue to orange, indicating siderophore production. **(D)** Magnified view of JC1 colonies from the plate in **(C)**, showing the orange halos in the media indicative of siderophore production. Scale bars, 1.5 cm **(A–C)**, 4 mm **(D)**.

australis by *Phaeobacter gallaeciensis* 2.10 (Rao et al., 2007). In other roseobacters, quorum sensing regulates secondary metabolite production, such as TDA in *Phaeobacter gallaeciensis* (Berger et al., 2011). In the indigoidine producing roseobacter, *Leisingera* sp. Y4I, there are two quorum sensing systems that regulate indigoidine production, *pgaIR* and *phaIR* (Cude et al., 2015). The JC1 *luxI* homolog has a 72% amino acid similarity to *pgaI* (RBY4I_1689) in Y4I, and the JC1 *luxR* homolog has an 81% amino acid similarity to *pgaR* (RBY4I_3631) in Y4I. The second set of *luxIR* homologs in Y4I, *phaIR* (RBY4I_3464 and RBY4I_1027), is not present in JC1. PgaI synthesizes the C8-HSL, produced by several proteobacteria, while PhaI synthesizes the $3OHC_{12:1}$-HSL, which may be species specific. JC1 lacking the *phaIR* system may reflect its divergence from *Leisingera* sp. Y4I. Further analyses will be needed to understand if indigoidine production in *Leisingera* sp. JC1 is regulated by quorum sensing.

Siderophore Production

Three separate siderophore biosynthesis gene clusters were detected in the genome, as described above, and production of iron chelators was confirmed by plating on CAS agar (**Figures 2C,D**). Appearance of an orange halo around colonies indicates that iron was sequestered from the chrome-azurol S dye in the media. Siderophores are high-affinity iron chelators, and can provide a growth advantage to cells in iron-limited environments, such as in seawater and in colonization of hosts. Although, the presence of siderophore biosynthesis genes in the genomes of currently sequenced roseobacter clade members is rare, 10 of the 12 previously sequenced *E. scolopes* ANG roseobacter symbionts did have the genes and/or demonstrate production of siderophores (Collins et al., 2015). Similar to the majority of the squid-associated roseobacter clade, the *Leisingera* sp. JC1 genome contains siderophore biosynthesis genes, indicating that siderophore production may play a role in the ANG symbiosis.

Indigoidine Biosynthesis Genes

The indigoidine biosynthesis gene cluster in *Leisingera* sp. JC1 contains all six biosynthesis genes previously described for *Leisingera* sp. Y4I (Cude et al., 2012) and shares a similar genome

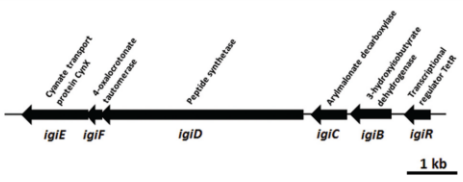


FIGURE 3 | *Leisingera* sp. JC1 indigoidine biosynthesis operon. Operon organization obtained from the published genome sequence (OBY26161.1, OBY26149.1, OBY26162.1, OBY26150.1, OBY26151.1, OBY26152.1). Arrows drawn to scale.

arrangement (**Figure 3**). To confirm the presence of individual members of the indigoidine biosynthesis gene cluster, primers were designed to three different components of the pathway, the non-ribosomal peptide synthetase (*igiD*), the transcriptional regulator (*igiR*), and one of the three indigoidine modification genes (*igiC*). The presence of these genes in JC1 genomic DNA was confirmed by PCR (Supplementary Figure S1).

Indigoidine biosynthesis genes have been detected in a diverse group of bacteria, including the *Actinobacteria* (*Streptomyces),* and *Alpha-, Beta-,* and *Gamma-proteobacteria*. The JC1 indigoidine biosynthesis operon shares the closest homology to the operon in *Leisingera* sp. Y4I, with 90–95% amino acid similarity for all gene products (**Table 2**). Other indigoidine biosynthesis operons share the non-ribosomal peptide synthetase, *igiD*, but many lack the same accessory genes required to modify indigoidine. When compared to other indigoidine producing strains, the *igiD* of JC1 is functionally homologous to other NRPS genes, sharing 49–53% amino acid similarity with *Vogesella indigofera*, *Streptomyces lavendulae*, and *Dickeya dadantii* 3937 (**Table 2**). A comparison with the genome of *Leisingera* sp. Y4I also confirmed that the indigoidine gene cluster is shared between these strains although absent from related ANG isolate *Leisingera* sp. M7 (Supplementary Figures S4 and S5).

Detection of Indigoidine Production by *Leisingera* sp. JC1

Because of the distinctive morphology and the genetic evidence for indigoidine biosynthesis, *Leisingera* sp. JC1 was cultured and extracted to obtain chemical evidence of indigoidine production. Using a three-step culture process, a deep blue liquid culture was obtained. However, upon extraction using a typical resin-based organic extraction protocol, most of the blue color was insoluble in organic solvents and little evidence of indigoidine production was observed via liquid chromatography-mass spectrometry (LC–MS, see **Figures 4D,E**), integrating to only 0.8% of the JC1 normal extract and indicative of negligible indigoidine extraction using this method. Therefore, an indigoidine enriched extraction protocol was utilized to pellet the insoluble indigoidine away from other media and cellular components followed by dissolving the sample in DMSO (Yu et al., 2013), resulting in an indigoidine enriched extract with 91.1% indigoidine. Analysis via LC–MS confirmed the presence of indigoidine (**Figure 4A**) in the indigoidine enriched extract (**Figures 4B,C**) with a peak eluting at 10.7 min with an $[M\text{-}H]^-$ of 247.0, consistent with the molecular weight and fragmentation pattern of indigoidine (248.2 g/mol) and in agreement with literature precedent (Yu et al., 2013). In addition, indigoidine was detected in two other JC and ANG isolates that exhibited a similar dark blue coloration in culture (data not shown).

Antibacterial Activity of *Leisingera* sp. JC1

Zone of Inhibition

Zone of inhibition assays were performed to test the ability of *Leisingera* sp. JC1 to inhibit other marine bacteria, both free-living and symbiotic (Supplementary Table S3). JC1 was tested against the *E. scolopes* light organ symbiont, *V. fischeri* ES114; another bioluminescent member of the *Vibrionaceae*, *P. leiognathi* KNH6, isolated from Hawaiian seawater; *V. harveyi* B392; *V. parahaemolyticus* KNH1 and *V. anguillarum* 775. These bacteria were plated at lawn densities from 10^4 to 10^7 CFU/mL to test the efficacy of possible inhibition at varying densities

TABLE 2 | Comparison of indigoidine biosynthesis operon in *Leisingera* sp. JC1 to other indigoidine producing strains.

Gene	Annotation	% Amino acid identity to *Leisingera* sp. Y4I operon	% Amino acid identity to *Vogesella indigofera* operon	% Amino acid identity to *Streptomyces lavendulae* operon	% Amino acid identity to *Dickeya dadantii* 3937 operon
igiE	Cyanate transport protein, CynX	95	58	NA	NA
igiF	4-oxalocrotonate tautomerase	91	NA*	NA	NA
igiD	Peptide synthetase	91	53	50	49
igiC	Arylmalonate decarboxylase	95	56	NA	NA
igiB	Hydroxyisobutyrate dehydrogenase	93	51	NA	NA
igiR	Transcriptional regulator, TetR	90	42	NA	NA

*NA, not applicable, no homolog of the gene present in that organism.

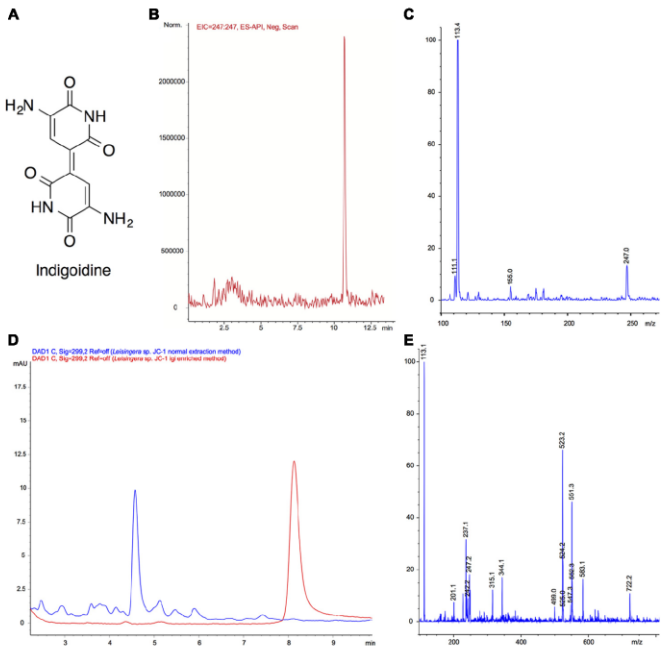


FIGURE 4 | Mass spectral confirmation of indigoidine production by *Leisingera* sp. JC1 and comparison of normal and igi enriched extracts. **(A)** Structure of indigoidine. **(B)** Extracted ion chromatogram (EIC) of indigoidine to confirm presence in extract ($[M\text{-}H]^-$ 247.0) using a previously reported LC–MS solvent gradient (Yu et al., 2013). **(C)** Negative ionization mass spectrum of peak at 10.7 min. The $[M\text{-}H]^-$ peak of 247.0 is consistent with a molecular formula of $C_{10}H_7N_4O_4$, confirming the presence of indigoidine (248.19 g/mol). **(D)** LC–MS UV chromatogram overlays (299 nm) of normal extract (blue) with indigoidine (igi) enriched extract (red). The peak eluting at t_R 8.2 min in the igi enriched extract was confirmed to be indigoidine with a purity of 91.1% (t_R different due to use of isocratic solvent system to minimize baseline variation). Although, a small peak was present at t_R 8.2 min in the normal extract, this integrates to only 0.8% indicating that the *Leisingera* sp. JC1 normal extract contains negligible amounts of indigoidine. **(E)** Negative ionization mass spectrum of peak at t_R 8.2 min for normal extraction method. Although there was an ion corresponding to $[M\text{-}H]^-$ of 247.2, this was not the most dominant ion and the more quantitative UV analysis in D shows only a very small peak, indicative of negligible indigoidine in the JC1 normal extract.

which more closely reflect biologically relevant concentrations. Overall, *Leisingera* sp. JC1 differentially inhibited the five vibrios tested (**Figure 5**). For two of the strains tested, *V. fischeri* ($F_{3,76} = 12.63$, $P < 0.0001$) and *P. leiognathi* ($F_{3,60} = 137.5$, $P < 0.0001$), JC1 showed significantly greater inhibition at lower lawn densities (**Figure 5A**; Supplementary Figure S2). When measured ZOIs were normalized for variations in JC1 colony diameter, there was an average 4.2 mm ZOI at 10^4 CFU/mL of *V. fischeri*, while at the 10^7 CFU/mL density, there was a 2.3 mm ZOI (Supplementary Figures S2I–L). A multiple comparisons *post hoc* Tukey test determined that the ZOI for the 10^4–10^6 CFU/mL lawn densities of *V. fischeri* were significantly greater than the ZOI at the 10^7 CFU/mL density. The change in ZOI with test strain lawn density was most apparent for *P. leiognathi*, where the average ZOI at the 10^4–10^5 CFU/mL lawn densities ranged from 2.4 to 2.9 mm, and then dropped to 0 mm at the 10^6–10^7 CFU/mL densities (**Figures 5A,D**; Supplementary Figures S2A–D). A multiple comparisons *post hoc* Tukey test showed that the ZOI at 10^4 CFU/mL of *P. leiognathi* was significantly different from the ZOI at 10^5 CFU/mL, and that both ZOIs at 10^4 and 10^5 CFU/mL were significantly different from the 10^6–10^7 CFU/mL results. *Leisingera* sp. JC1 showed a trend toward inhibition of *V. anguillarum* with ZOIs ranging from 2.7 to 3.3 mm (**Figures 5A,E**; Supplementary Figures S2E–H) although, a one-way ANOVA determined that the ZOIs were not statistically different ($F_{3,60} = 2.553$, $P = 0.0639$). No inhibition was observed when JC1 was tested against *V. parahaemolyticus* or *V. harveyi* at any lawn density (**Figures 5A,F,G**; Supplementary Figures S2M–P).

Leisingera sp. JC1 was also tested in a ZOI assay against the 12 previously described ANG isolates (Collins et al., 2015) and one additional ANG isolate, *Muricauda* sp. ANG21. All ANG isolates were only tested at a lawn density of approximately 10^8 CFU/mL. Inhibition was observed against *Ruegeria* sp. ANG-S4,

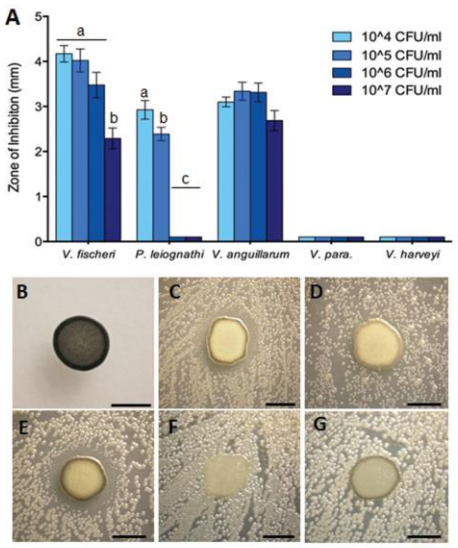


FIGURE 5 | *Leisingera* sp. JC1 differentially inhibits pathogenic and non-pathogenic vibrio species. (A) Zones of inhibition around JC1 colonies were measured with increasing lawn densities of each vibrio species. *Leisingera* sp. JC1 differentially inhibited five vibrio strains. Significantly greater inhibition of *Vibrio fischeri* was seen between the 10^4–10^6 CFU/ml and the 10^7 CFU/ml lawn density. Significantly greater inhibition of *Photobacterium leiognathi* was seen between the 10^4 and 10^5 CFU/ml lawn densities, as well as between the 10^4–10^5 and the 10^6–10^7 CFU/ml densities. JC1 inhibited *Vibrio anguillarum* uniformly across all lawn densities and did not inhibit *Vibrio parahaemolyticus* or *Vibrio harveyi*. Letters indicate significantly different groups based on results from *post hoc* Tukey tests (see Results and Discussion). Representative 24 h images of **(B)** JC1 plated alone on YTSS agar and with 10^6 CFU/ml of **(C)** *V. fischeri*, **(D)** *P. leiognathi*, **(E)** *V. anguillarum*, **(F)** *V. parahaemolyticus*, and **(G)** *V. harveyi*. Scale bars: 5 mm **(B–G)**.

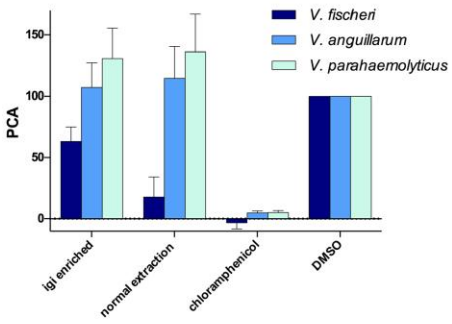


FIGURE 6 | *Leisingera* sp. JC1 extracts inhibit *V. fischeri* in 96-well liquid assays. JC1 cultures were grown at large scale and extracts were prepared using either an indigoidine enriched procedure (igi enriched) or a normal extraction method. All extracts were screened at 500 μg/mL against *V. fischeri*, *V. anguillarum*, and *V. parahaemolyticus*. The JC1 normal extract strongly inhibited growth of *V. fischeri* (PCA 17.9 ± 9.3) with less activity of the JC1 igi enriched extract against *V. fischeri* (PCA 63.3 ± 6.7). No inhibition of *V. anguillarum* or *V. parahaemolyticus* was observed. Data shown are average PCA values from three experiments (PCA, percent control activity, as compared with DMSO negative control).

with an average ZOI of 6.3 mm (±0.7) and against *Muricauda* sp. ANG21, with an average ZOI of 5.9 mm (±0.6; Supplementary Figure S3). *Leisingera* sp. JC1 was not able to inhibit any of the other *Leisingera* spp. previously isolated from ANGs. Since partitioning between bacterial taxa is observed in the ANG tubules some activity against other ANG isolates may contribute to competition between strains during colonization (Collins et al., 2012).

96-Well Liquid Assay

Both the normal and indigoidine enriched JC1 extracts were screened for activity using 96-well plate liquid assays with several of the vibrios tested above, including *V. fischeri* ES114, *V. anguillarum* 775, and *V. parahaemolyticus* KNH1 (**Figure 6**). Both extracts were initially tested at 500 μg/mL with minimum inhibitory concentrations (MICs) determined for active samples. The JC1 normal extract was found to strongly inhibit growth of *V. fischeri* with a PCA value of 17.9 ± 9.3 during screening and was determined to have a MIC of 250 μg/mL. The JC1 indigoidine enriched extract also exhibited moderate inhibition of *V. fischeri* with a PCA of 63.3 ± 6.7.

In contrast to the ZOI data above, no inhibition was observed for either extract when tested against *V. anguillarum*, potentially due to differences between the activity of indigoidine in agar versus liquid assays, as seen with *Leisingera* sp. Y41 and hypothesized to result from changes in the redox state of indigoidine (Cude et al., 2012). These results may also be attributed to differences in the chemical composition between extracts and the bacteria *in situ* (e.g., aqueous soluble metabolites are generally excluded from the extraction protocols used in this study). Neither JC1 extract inhibited *V. parahaemolyticus*, in agreement with the ZOI data above.

Previous studies with a mutant of *Leisingera* sp. Y4I that did not produce indigoidine suggested that production of the compound is required for inhibition of *V. fischeri* (Cude et al., 2012). However, with the more potent inhibition of *V. fischeri* seen in the JC1 normal extract versus the indigoidine enriched extract in this study (**Figure 6**), indigoidine production does not seem to be the only mechanism of inhibition for *Leisingera* sp. JC1. Given that the JC1 normal extract contains only minimal amounts of indigoidine (0.8% as discussed above), the bacterium may be utilizing other secondary metabolites in conjunction with indigoidine for chemical defense. The JC1 genome includes several other secondary metabolite biosynthetic gene clusters for HSL, siderophore, bacteriocin, PKS, and PKS/NRPS production and thus *Leisingera* sp. JC1 likely utilizes one or more of the

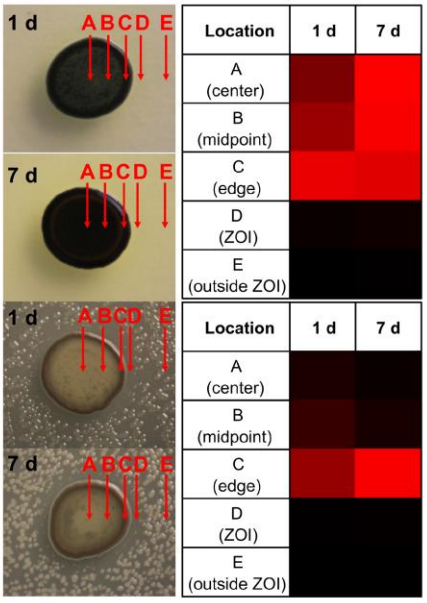


FIGURE 7 | *Leisingera* sp. JC1 localizes indigoidine production to outer edge of colony in the presence of *V. fischeri*. Five locations (A–E) were analyzed for the presence of indigoidine via direct analysis in real time-mass spectrometry (DART-MS) over 7 days. (A) In the absence of *V. fischeri*, indigoidine production was more uniformly distributed across the colony. (B) In the presence of *V. fischeri*, indigoidine production was primarily localized to outer edge of colony (location C) whereas minimal indigoidine was detected in the center of the colony (location A). Very little indigoidine was detected in the ZOI or outside ZOI in the presence or absence of *V. fischeri*. Heatmaps represent a gradient of indigoidine ion abundance from less abundance (black, 0%) to more abundance (red, 100%).

compounds encoded by these pathways for chemical defense, in addition to the defensive capabilities attributed to indigoidine. Creating an indigoidine mutant of *Leisingera* sp. JC1 will help test this hypothesis, in conjunction with identification of additional metabolite(s) responsible for JC1 antimicrobial activity.

Localization of Indigoidine Production by *Leisingera* sp. JC1

While performing ZOI assays, there was a dramatic change in colony pigmentation of *Leisingera* sp. JC1 when grown alone (**Figure 5B**) as compared to growth under challenge with various vibrio strains (**Figures 5C–G**). Deep blue pigment production was observed uniformly when JC1 was grown in monoculture and appeared to localize to the outer edges of the colonies when presented with vibrio strains. Direct analysis in real time-mass spectrometry (DART-MS) is an ambient ionization technique in which samples can be analyzed without sample preparation or extraction (Sanchez et al., 2011). DART-MS was utilized to chemically confirm the visual observations of localization of indigoidine production of JC1 in monoculture and co-culture with *V. fischeri* over the course of 7 days (**Figure 7**). Five locations were selected on each colony including center (A), midpoint (B), edge (C), ZOI (D), and outside ZOI (E). In the absence of *V. fischeri*, indigoidine was uniformly produced throughout JC1 colonies (locations A–C). However, in the presence of *V. fischeri* there was little to no indigoidine production in the center or midpoints of JC1 colonies, but intense indigoidine detected around colony edges (**Figure 7**, location C only). Indigoidine was only minimally detected in the ZOI or outside the ZOI for either monoculture or co-culture. Trends in the localization of indigoidine production were even more apparent upon measurement after 7 days.

There are several examples of pigment production being induced when in the presence of other bacteria, such as in *Staphylococcus aureus* when co-cultured with *Pseudomonas aeruginosa* (Antonic et al., 2013) or production of a red pigment by *Streptomyces lividans* TK23 when co-cultured with *Tsukamurella pulmonis* TP-B0596 (Onaka et al., 2011). Pigment production can also be induced under other stress response conditions, such as protection from UV radiation (Tong and Lighthart, 1997). Pigment production has also been tied to photosynthesis (Orf and Blankenship, 2013), however, *Leisingera* sp. JC1 lacks genes associated with photosynthesis or carbon fixation (data not shown).

When grown alone, *Leisingera* sp. JC1 exhibited a uniform blue–black pigmentation across the colony which was confirmed by mass spectrometry to be essentially uniform production of indigoidine. Secondary metabolite biosynthesis is an energy intensive endeavor and production of antimicrobial compounds would typically be thought to be reserved for defensive situations. Since indigoidine is produced throughout the colony when in monoculture, and given its relatively moderate antibacterial activity as suggested by assays with the indigoidine enriched extract, it is also possible that indigoidine serves multiple functions for *Leisingera* sp. JC1. However, *Leisingera* sp. JC1 localized indigoidine production to the outer edges of the colony when co-cultured with *V. fischeri* and other vibrios. If utilized as a defensive compound, indigoidine may be localized to points of direct interaction with other microorganisms. Secondary metabolite production can be localized to susceptible locations such as in plants, sponges, and other sessile terrestrial and marine organisms (Amsler et al., 2001; Furrow et al., 2003; Van Dyck et al., 2010). The role of *Leisingera* sp. JC1 has yet to be examined directly in the ANG symbiosis but localized production of indigoidine or other secondary metabolites may play a role in egg defense or inhibition of other bacteria from colonizing the ANG (see conclusions below).

Regulation of Indigoidine Production by *Leisingera* sp. JC1

After observing localized production of indigoidine when grown on solid media with *V. fischeri*, additional co-culture experiments were undertaken in liquid media using several

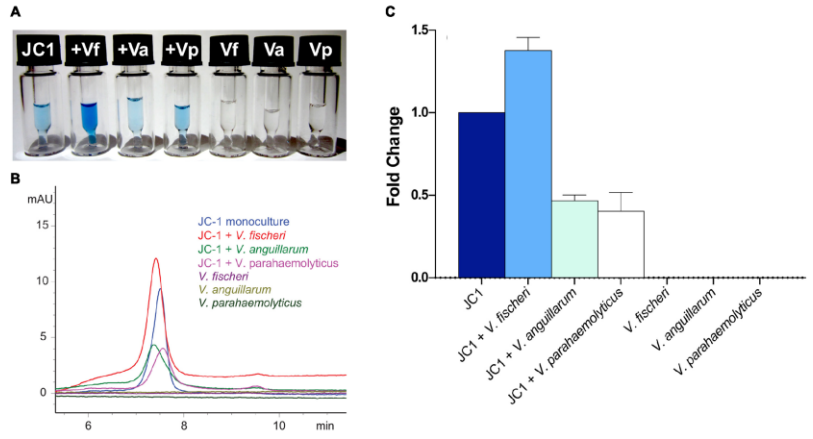


FIGURE 8 | *Leisingera* sp. JC1 upregulates indigoidine production in the presence of *V. fischeri*. JC1 was cultured alone and in the presence of *V. fischeri*, *V. anguillarum*, and *V. parahaemolyticus* (each vibrio was also monocultured). **(A)** Extracts were prepared in DMSO at 5 mg/mL for LC–MS (from left to right): JC1 monoculture, JC1 + *V. fischeri*, JC1 + *V. anguillarum*, JC1 + *V. parahaemolyticus*, *V. fischeri* monoculture, *V. anguillarum* monoculture, and *V. parahaemolyticus* monoculture. As shown, the highest pigment production was observed when JC1 was co-cultured with *V. fischeri*. **(B)** LC–MS UV chromatograms (299 nm) for all extracts. JC1 indigoidine production was upregulated when co-cultured with *V. fischeri* (red). Down regulation was observed when co-cultured with *V. anguillarum* or *V. parahaemolyticus* (green and pink, respectively). **(C)** Following extraction, indigoidine production was measured using HPLC and quantitated via measurement of the area under the curve for UV absorption at 299 nm. Fold change was calculated as compared with JC1 monoculture.

of the vibrios from the antimicrobial assays above. *Leisingera* sp. JC1 was grown in monoculture and in the presence of *V. fischeri*, *V. anguillarum*, and *V. parahaemolyticus*, followed by extraction and measurement of indigoidine production (**Figure 8**). Monocultures of all three vibrios were also grown and extracted as controls. Addition of *V. fischeri* to established cultures of *Leisingera* sp. JC1 resulted in a 1.38 fold increase in indigoidine. Co-cultures with *V. anguillarum* and *V. parahaemolyticus* resulted in a decrease in indigoidine production of approximately 0.5 fold for both organisms. Vibrio monocultures confirmed that these species do not produce indigoidine. Changes in indigoidine production were also visually evident with darker, more intense blue observed for extracts cultured with *V. fischeri* in comparison with JC1 monoculture, as well as lighter blue extracts observed for *V. anguillarum* and *V. parahaemolyticus* co-cultures (**Figure 8A**).

The increase in indigoidine production of JC1 with *V. fischeri* is consistent with the antibacterial activity observed for *Leisingera* sp. JC1 on both solid and liquid media (**Figures 5** and **6**), strengthening the hypothesis that indigoidine may play a protective role in association with *E. scolopes*. In addition, the downregulation of production with *V. anguillarum* and *V. parahaemolyticus* also supports the liquid culture bioassay data (**Figure 6**). The differential antimicrobial activity and indigoidine production between the three vibrios may be due to the purported role of the ANG and JC bacteria in the host. The ability of *Leisingera* sp. JC1 to inhibit *V. fischeri* may be related to the fact that the ANG is located directly posterior to the light organ, which harbors high densities of the sole symbiont, *V. fischeri* (McFall-Ngai, 2014). Each day 95% of viable *V. fischeri* cells in the light organ are expelled directly into the mantle cavity of the host as part of the regulatory mechanisms of that association (Boettcher et al., 1996; Nyholm and McFall-Ngai, 1998). A study from another squid, *Doryteuthis pealeii* (Kaufman et al., 1998) suggests that ANG bacteria are environmentally transmitted during development. Given that *V. fischeri* is not detected in the ANG (Collins et al., 2012), the inhibitory effect of *Leisingera* sp. JC1 and other ANG isolates may prevent *V. fischeri* and other vibrios from colonizing the ANG and thus help shape the consortium during development. Alternatively, inhibition against vibrios may play a role in egg defense since eggs are exposed to seawater for approximately three weeks and vibrios are known to be common members of the bacterioplankton.

CONCLUSIONS

Genome analyses confirm that *Leisingera* sp. JC1 is part of the squid-associated roseobacter clade. Both *in silico* and *in vitro* analyses confirmed the secondary metabolite potential and production of siderophores, acyl-homoserine lactones associated with quorum sensing, and the pigment indigoidine. *Leisingera* sp. JC1 and its extracts had inhibitory activity against a variety of marine bacteria including the light organ symbiont *V. fischeri*. Furthermore, JC1 challenged with *V. fischeri* led to increased localized

production of indigoidine as well as an increased production of indigoidine when co-cultured in liquid media. Taken together these results suggest that *Leisingera* sp. JC1 may play a protective role in egg defense and/or in shaping the microbial community of the ANG. The importance of defensive symbioses in nature is becoming increasingly more evident (Flórez et al., 2015). A number of both terrestrial and marine organisms use novel secondary metabolites produced by bacteria toward defense from potential pathogens and fouling microorganisms. Since roseobacters have been found in the ANGs of a number of cephalopods from diverse marine environments (Kaufman et al., 1998; Grigioni et al., 2000; Barbieri et al., 2001; Pichon et al., 2005; Collins et al., 2012) there may be a conserved function of this group in this symbiosis. Further studies from this group may reveal novel compounds that are important for the biology of these associations and that exhibit antimicrobial activity.

AUTHOR CONTRIBUTIONS

MB, SN, SG, and AS conceptualized and designed research; SG, AS, MF, and JLG conducted experiments; MB, SN, SG, AS, MF, and JPG analyzed data and wrote the paper.

FUNDING

This research was funded by NSF IOS-1557914 to SN and MB, University of Connecticut Office of the Vice President for Research to SN, and the University of Connecticut Outstanding Multicultural Scholars Program to AS.

ACKNOWLEDGMENTS

The authors would like to thank Anne A. Sung for assistance with CFU counts, Alison Buchan for donation of *Leisingera* sp. Y4I, Allison H. Kerwin for helpful comments, and Kewalo Marine Laboratory of the University of Hawaii for assistance with animal collections.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2016.01342

REFERENCES

Alikhan, N.-F., Petty, N. K., Ben Zakour, N. L., and Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12:402. doi: 10.1186/1471-2164-12-402

Amsler, C. D., McClintock, J. B., and Baker, B. J. (2001). Secondary metabolites as mediators of trophic interactions among Antarctic marine organisms. *Am. Zool.* 41, 17–26.

Antonic, V., Stojadinovic, A., Zhang, B., Izadjoo, M. J., and Alavi, M. (2013). *Pseudomonas* aeruginosa induces pigment production and enhances virulence in a white phenotypic variant of *Staphylococcus aureus*. *Infect. Drug Resist.* 6, 175–186. doi: 10.2147/IDR.S49039

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75

Barbieri, E., Barry, K., Child, A., and Wainwright, N. (1997). Antimicrobial activity in the microbial community of the accessory nidamental gland and egg cases of Loligo pealei (Cephalopoda: Loliginidae). *Biol. Bull.* 193, 275–276.

Barbieri, E., Paster, B., Hughes, D., Zurek, L., Moser, D., Teske, A., et al. (2001). Phylogenetic characterization of epibiotic bacteria in the accessory nidamental gland and egg capsules of the squid Loligo pealei (Cephalopoda:Loliginidae). *Environ. Microbiol.* 3, 151–167. doi: 10.1046/j.1462-2920.2001.00172.x

Berger, M., Neumann, A., Schulz, S., Simon, M., and Brinkhoff, T. (2011). Tropodithietic acid production in *Phaeobacter gallaeciensis* is regulated by N-acyl homoserine lactone-mediated quorum sensing. *J. Bacteriol.* 193, 6576–6585. doi: 10.1128/JB.05818-11

Boettcher, K. J., Ruby, E. G., and McFall-Ngai, M. J. (1996). Bioluminescence in the symbiotic squid *Euprymna scolopes* is controlled by a daily biological rhythm. *J. Comp. Physiol.* 179, 65–73. doi: 10.1007/BF00193435

Bruhn, J. B., Gram, L., and Belas, R. (2007). Production of antibacterial compounds and biofilm formation by *Roseobacter* species are influenced by culture conditions. *Appl. Environ. Microbiol.* 73, 442–450. doi: 10.1128/AEM.02238-06

Bruhn, J. B., Haagensen, J. A. J., Bagge-Ravn, D., and Gram, L. (2006). Culture conditions of *Roseobacter* strain 27-4 affect its attachment and biofilm formation as quantified by real-time PCR. *Appl. Environ. Microbiol.* 72, 3011–3015. doi: 10.1128/AEM.72.4.3011-3015.2006

Cha, C., Gao, P., Chen, Y. C., Shaw, P. D., and Farrand, S. K. (1998). Production of acyl-homoserine lactone quorum-sensing signals by gram-negative plant-associated bacteria. *Mol. Plant Microbe Interact.* 11, 1119–1129. doi: 10.1094/MPMI.1998.11.11.1119

Chilton, M. D., Currier, T. C., Farrand, S. K., Bendich, A. J., Gordon, M. P., and Nester, E. W. (1974). *Agrobacterium tumefaciens* DNA and PS8 bacteriophage DNA not detected in crown gall tumors. *Proc. Natl. Acad. Sci. U.S.A.* 71, 3672–3676. doi: 10.1073/pnas.71.9.3672

Cianfanelli, F. R., Monlezun, L., and Coulthurst, S. J. (2016). Aim, load, and fire: the Type VI secretion system, a bacterial nanoweapon. *Trends Microbiol.* 24, 51–62. doi: 10.1016/j.tim.2015.10.005

Clinical and Laboratory Standards Institute (2012). *Methods for Dilution Antimicrobial Susceptibility Tests for Bacteria that Grow Aerobically. M07A9*, 9th Edn. Wayne, PA: Clinical and Laboratory Standards Institute.

Collins, A. J., Fullmer, M. S., Gogarten, J. P., and Nyholm, S. V. (2015). Comparative genomics of *Roseobacter* clade bacteria isolated from the accessory nidamental gland of *Euprymna scolopes*. *Front. Microbiol.* 6:123. doi: 10.3389/fmicb.2015.00123

Collins, A. J., LaBarre, B. A., Won, B. S., Shah, M. V., Heng, S., Choudhury, M. H., et al. (2012). Diversity and partitioning of bacterial populations within the accessory nidamental gland of the squid *Euprymna scolopes*. *Appl. Environ. Microbiol.* 78, 4200–4208. doi: 10.1128/AEM.07437-11

Collins, A. J., and Nyholm, S. V. (2011). Draft genome of *Phaeobacter gallaeciensis* ANG1, a dominant member of the accessory nidamental gland of *Euprymna scolopes*. *J. Bacteriol.* 193, 3397–3398. doi: 10.1128/JB.05139-11

Cude, W. N., Mooney, J., Tavanaei, A. A., Hadden, M. K., Frank, A. M., Gulvik, C. A., et al. (2012). Production of the antimicrobial secondary metabolite indigoidine contributes to competitive surface colonization by the marine *Roseobacter Phaeobacter* sp. strain Y4I. *Appl. Environ. Microbiol.* 78, 4771–4780. doi: 10.1128/AEM.00297-12

Cude, W. N., Prevatte, C. W., Hadden, M. K., May, A. L., Smith, R. T., Swain, C. L., et al. (2015). Phaeobacter sp. strain Y4I utilizes two separate cell-to-cell communication systems to regulate production of the antimicrobial indigoidine. *Appl. Environ. Microbiol.* 81, 1417–1425. doi: 10.1128/AEM.02551-14

Darling, A. E., Mau, B., and Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* 5:e11147. doi: 10.1371/journal.pone.0011147

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* 9, 772. doi: 10.1038/nmeth.2109

Durand, E., Cambillau, C., Cascales, E., and Journet, L. (2014). VgrG, Tae, Tle, and beyond: the versatile arsenal of Type VI secretion effectors. *Trends Microbiol.* 22, 498–507. doi: 10.1016/j.tim.2014.06.004

Edenborn, H. M., and Litchfield, C. D. (1985). Glycolate metabolism by *Pseudomonas* sp., *strain S227*, isolated from a coastal marine sediment. *Mar. Bio.* 88, 199–205. doi: 10.1007/BF00397167

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Flórez, L. V., Biedermann, P. H., Engl, T., and Kaltenpoth, M. (2015). Defensive symbioses of animals with prokaryotic and eukaryotic microorganisms. *Nat. Prod. Rep.* 32, 904–936. doi: 10.1039/c5np00010f

Furrow, F. B., Amsler, C. D., McClintock, J. B., and Baker, B. J. (2003). Surface sequestration of chemical feeding deterrents in the Antarctic sponge Latrunculia apicalis as an optimal defense against sea star spongivory. *Mar. Bio.* 143, 443–449. doi: 10.1007/s00227-003-1109-5

Geng, H., Bruhn, J. B., Nielsen, K. F., Gram, L., and Belas, R. (2008). Genetic dissection of tropodithietic acid biosynthesis by marine roseobacters. *Appl. Environ. Microbiol.* 74, 1535–1545. doi: 10.1128/AEM.02339-07

Grigioni, S., Boucher-Rodoni, R., Demarta, A., Tonolla, M., and Peduzzi, R. (2000). Phylogenetic characterisation of bacterial symbionts in the accessory nidamental glands of the sepioid *Sepia officinalis* (Cephalopoda: Decapoda). *Mar. Biol.* 136, 217–222. doi: 10.1007/s002270050679

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Hood, R. D., Singh, P., Hsu, F., Güvener, T., Carl, M. A., Trinidad, R. R. S., et al. (2010). A Type VI secretion system of *Pseudomonas aeruginosa* targets a toxin to bacteria. *Cell Host Microbe* 7, 25–37. doi: 10.1016/j.chom.2009.12.007

Jiang, F., Waterfield, N. R., Yang, J., Yang, G., and Jin, Q. (2014). A *Pseudomonas aeruginosa* type VI secretion phospholipase D effector targets both prokaryotic and eukaryotic cells. *Cell Host Microbe* 15, 600–610. doi: 10.1016/j.chom.2014.04.010

Kaufman, M., Ikeda, Y., Patton, C., van Dykhuizen, G., and Epel, D. (1998). Bacterial symbionts colonize the accessory nidamental gland of the squid Loligo opalescens via horizontal transmission. *Biol. Bull.* 194, 36–43. doi: 10.2307/1542511

Martens, T., Gram, L., Grossart, H. P., Kessler, D., Müller, R., Simon, M., et al. (2007). Bacteria of the *Roseobacter* clase show potential for secondary metabolite production. *Microb. Ecol.* 54, 31–42. doi: 10.1007/s00248-006-9165-2

McFall-Ngai, M. J. (2014). The importance of microbes in animal development: lessons from the squid-vibrio symbiosis. *Ann. Rev. Microbiol.* 68, 177–194. doi: 10.1146/annurev-micro-091313-103654

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14:60. doi: 10.1186/1471-2105-14-60

Newton, R. J., Griffin, L. E., Bowles, K. M., Meile, C., Gifford, S., Givens, C. E., et al. (2010). Genome characteristics of a generalist marine bacterial lineage. *ISME J.* 4, 784–798. doi: 10.1038/ismej.2009.150

Nyholm, S. V., and McFall-Ngai, M. J. (1998). Sampling the light-organ microenvironment of *Euprymna scolopes*: description of a population of host cells in association with the bacterial symbiont Vibrio fischeri. *Biol. Bull.* 195, 89–97. doi: 10.2307/1542815

Onaka, H., Mori, Y., Igarashi, Y., and Furumai, T. (2011). Mycolic acid-containing bacteria induce natural-product biosynthesis in *Streptomyces* species. *Appl. Environ. Microbiol.* 77, 400–406. doi: 10.1128/AEM.01337-10

Orf, G. S., and Blankenship, R. E. (2013). Chlorosome antenna complexes from green photosynthetic bacteria. *Photosynth. Res.* 116, 315–331. doi: 10.1007/s11120-013-9869-3

Pichon, D., Gaia, V., Norman, M. D., and Boucher-Rodoni, R. (2005). Phylogenetic diversity of epibiotic bacteria in the accessory nidamental glands of squids (Cephalopoda: Loliginidae and *Idiosepiidae*). *Mar. Biol.* 147, 1323–1332. doi: 10.1007/s00227-005-0014-5

Pukatzki, S., Ma, A. T., Sturtevant, D., Krastins, B., Sarracino, D., Nelson, W. C., et al. (2006). Identification of a conserved bacterial protein secretion system in *Vibrio cholerae* using the *Dictyostelium* host model system. *Proc. Natl. Acad. Sci. U.S.A.* 103, 1528–1533. doi: 10.1073/pnas.0510322103

Rao, D., Webb, J. S., Holmström, C., Case, R., Low, A., Steinberg, P., et al. (2007). Low densities of epiphytic bacteria from the marine alga Ulva australis inhibit settlement of fouling organisms. *Appl. Environ. Microbiol.* 73, 7844–7852. doi: 10.1128/AEM.01543-07

Ravn, L., Christensen, A. B., Molin, S., Givskov, M., and Gram, L. (2001). Methods for detecting acylated homoserine lactones produced by Gram-negative bacteria and their application in studies of AHL-production kinetics. *J. Microbiol. Methods* 44, 239–251. doi: 10.1016/S0167-7012(01)00217-2

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends Genet.* 16, 276–277. doi: 10.1016/S0168-9525(00)02024-2

Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U.S.A.* 106, 19126–19131. doi: 10.1073/pnas.0906412106

Ronquist, F., Teslenko, M., Mark, P., van der Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Ruiz, B., Chávez, A., Forero, A., García-Huante, Y., Romero, A., Sánchez, M., et al. (2010). Production of microbial secondary metabolites: regulation by the carbon source. *Crit. Rev. Microbiol.* 36, 146–167. doi: 10.3109/10408410903489576

Sana, T. G., Hachani, A., Bucior, I., Soscia, C., Garvis, S., Termine, E., et al. (2012). The second type VI secretion system of *Pseudomonas aeruginosa* strain PAO1 is regulated by quorum sensing and fur and modulates internalization in epithelial cells. *J. Biol. Chem.* 287, 27095–27105. doi: 10.1074/jbc.M112.376368

Sanchez, L. M., Curtis, M. E., Bracamonte, B. E., Kurita, K. L., Navarro, G., Sparkman, O. D., et al. (2011). Versatile method for the detection of covalently bound substrates on solid supports by DART mass spectrometry. *Org. Lett.* 13, 3770–3773. doi: 10.1021/ol201404v

Schleicher, T., and Nyholm, S. (2011). Characterizing the host and symbiont proteomes in the association between the Bobtail squid, *Euprymna scolopes*, and the bacterium, *Vibrio fischeri*. *PLoS ONE* 6:e25649. doi: 10.1371/journal.pone.0025649

Schmidt, E. W., and Donia, M. S. (2010). Life in cellulose houses: symbiotic bacterial biosynthesis of ascidian drugs and drug leads. *Curr. Opin. Biotechnol.* 21, 827–833. doi: 10.1016/j.copbio.2010.10.006

Schneider, C. A., Rasband, W. S., and Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* 9, 671–675. doi: 10.1038/nmeth.2089

Schwarz, S., Hood, R. D., and Mougous, J. D. (2010). What is type VI secretion doing in all those bugs? *Trends Microbiol.* 18, 531–537. doi: 10.1016/j.tim.2010.09.001

Seyedsayamdost, M. R., Case, R. J., Kolter, R., and Clardy, J. (2011). The Jekyll-and-Hyde chemistry of *Phaeobacter gallaeciensis*. *Nat. Chem.* 3, 331–335. doi: 10.1038/nchem.1002

Tong, Y. Y., and Lighthart, B. (1997). Solar radiation is shown to select for pigmented bacteria in the ambient outdoor atmosphere. *Photochem. Photobiol.* 65, 103–106. doi: 10.1111/j.1751-1097.1997.tb01884.x

Tritt, A., Eisen, J. A., Facciotti, M. T., and Darling, A. E. (2012). An integrated pipeline for de novo assembly of microbial genomes. *PLoS ONE* 7:e42304. doi: 10.1371/journal.pone.0042304

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., et al. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Res.* 40, 1–12. doi: 10.1093/nar/gks596

Unterweger, D., Miyata, S. T., Bachmann, V., Brooks, T. M., Mullins, T., Kostiuk, B., et al. (2014). The *Vibrio cholerae* type VI secretion system employs diverse effector modules for intraspecific competition. *Nat. Comm.* 5, 3549. doi: 10.1038/ncomms4549

Van Dyck, S., Flammang, P., Meriaux, C., Bonnel, D., Salzet, M., Fournier, I., et al. (2010). Localization of secondary metabolites in marine invertebrates: contribution of MALDI MSI for the study of saponins in Cuvierian tubules of H. *forskali*. *PLoS ONE* 5:e13923. doi: 10.1371/journal.pone.0013923

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., et al. (2015). antiSMASH 3.0 - a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* 43, W237–W243. doi: 10.1093/nar/gkv437

Whistler, C. A., and Ruby, E. G. (2003). GacA regulates symbiotic colonization traits of *Vibrio fischeri* and facilitates a beneficial association with an animal host. *J. Bacteriol.* 185, 7202–7212. doi: 10.1128/JB.185.24.7202

Yu, D., Xu, F., Valiente, J., Wang, S., and Zhan, J. (2013). An indigoidine biosynthetic gene cluster from *Streptomyces chromofuscus* ATCC 49982 contains an unusual IndB homologue. *J. Ind. Microbiol. Biotechnol.* 40, 159–168. doi: 10.1007/s10295-012-1207-9

Zan, J., Cicirelli, E. M., Mohamed, N. M., Sibhatu, H., Kroll, S., Choi, O., et al. (2012). A complex LuxR-LuxI type quorum sensing network in a roseobacterial marine sponge symbiont activates flagellar motility and inhibits biofilm formation. *Mol. Microbiol.* 85, 916–933. doi: 10.1111/j.1365-2958.2012.08149.x

Zgoda, J. R., and Porter, J. R. (2001). A convenient microdilution method for screening natural products against bacteria and fungi. *Pharm. Biol.* 39, 221–225. doi: 10.1076/phbi.39.3.221.5934

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.